ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
ΚΑΤΕΥΘΥΝΣΗ: ΕΦΑΡΜΟΣΜΕΝΑ ΜΑΘΗΜΑΤΙΚΑ ΣΤΙΣ ΕΠΙΣΤΗΜΕΣ ΜΗΧΑΝΙΚΩΝ

ΘΕΜΑ

ΓΡΑΜΜΙΚΗ ΑΛΓΕΒΡΑ ΚΑΙ ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ

Εκπονούσα : **ΣΑΡΙΚΑΚΗ ΑΡΓΥΡΩ**

Επιβλέπων καθηγητής : **ΤΡΥΦΩΝ ΔΑΡΑΣ**

Χανιά, Μάρτιος 2018

ΕΥΧΑΡΙΣΤΙΕΣ .

Θα ήθελα να εκφράσω τις θερμές ευχαριστίες μου στον επιβλέπωντα καθηγητή μου Τρύφωνα Δάρα για την πολύτιμη καθοδήγηση του και την άψογη συνεργασία του καθ' όλη τη διάρκεια εκπόνησης της μεταπτυχιακής μου διατριβής.

Εν συνεχεία, θα ήθελα να ευχαριστήσω τους καθηγητές Πετράκη Μίνωα και Μαρινάκη Ιωάννη, όπου ήταν μέλη τις τριμελής επιτροπής μου.

Και τέλος ένα μεγάλο ευχαριστώ στην οικογένεια και τους φίλους μου, που με στηρίζουν σε κάθε βήμα μου.

Περίληψη

ΠΟΛΥΔΙΑΣΤΑΤΑ ΔΕΔΟΜΕΝΑ

Τα δεδομένα που καλούμαστε να επεξεργαστούμε κατά πλειοψηφία είναι πολυδιάστατα. Η ανάλυση πολυδιάστατων δεδομένων είναι ιδιαίτερα δύσκολη και η επίλυση του προβλήματος αυτού επιτυγχάνεται με τεχνικές μείωσης των διαστάσεων.

ΟΦΕΛΗ ΜΕΙΩΣΗΣ ΠΟΛΛΩΝ ΔΙΑΣΤΑΣΕΩΝ

Υπάρχει ποικιλία οφελών από τη μείωση των πολλών διαστάσεων.
Μερικά από αυτά είναι:

- Η μείωση των δεδομένων μπορεί να μας οδηγήσει σε πιο κατανοητά μοντέλα.
- Πολλοί αλγόριθμοι εξόρυξης δεδομένων λειτουργούν καλύτερα όταν το πλήθος διαστάσεων είναι μικρότερο.
- Σύμφωνα με 'την κατάρα των πολλών διαστάσεων' όσο μεγαλύτερο είναι το πλήθος διαστάσεων τα δεδομένα γίνονται πολύ αραιά στον χώρο που καταλαμβάνουν άρα επιδιώκουμε την μείωση τους.

ΤΕΧΝΙΚΕΣ ΜΕΙΩΣΗΣ ΤΩΝ ΔΙΑΣΤΑΣΕΩΝ

Οι τεχνικές μείωσης των διαστάσεων κατηγοριοποιούνται σε γραμμικές και μη γραμμικές. Με τον όρο γραμμικές εννοούμε τις μεθόδους αυτές που χρησιμοποιούν τεχνικές της Γραμμικής Άλγεβρας όπως η μελέτη διανυσμάτων, διανυσματικών χώρων, γραμμικών απεικονίσεων και συστημάτων γραμμικών εξισώσεων για να προβάλλουν δεδομένα από ένα χώρο πολλαπλών διαστάσεων σε έναν χώρο μικρότερου πλήθους διαστάσεων. Μια τέτοια γραμμική τεχνική μείωσης διαστάσεων αποτελεί και η λεγόμενη ΑΝΑΛΥΣΗ ΣΕ ΚΥΡΙΕΣ ΣΥΝΙΣΤΩΣΕΣ.

ΤΕΧΝΙΚΗ ΑΝΑΛΥΣΗΣ ΣΕ ΚΥΡΙΕΣ ΣΥΝΙΣΤΩΣΕΣ

Αποτελεί μια από τις δημοφιλείς γραμμικές μεθόδους ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ για τη μείωση διαστάσεων. Η Ανάλυση Κυρίων Συνιστωσών (Α.Κ.Σ) είναι μια στατιστική διαδικασία η οποία μετατρέπει μια αρχική ομάδα παρατηρήσεων δυνητικά συσχετιζόμενων μεταβλητών σε μια ομάδα νέων παρατηρήσεων με μεταβλητές οι οποίες είναι γραμμικός συνδυασμός των αρχικών έτσι ώστε να είναι ασυσχέτιστες μεταξύ τους και να περιέχουν όσο το δυνατόν μεγαλύτερο μέρος των διακυμάνσεων των αρχικών μεταβλητών. Καταλήγουμε δηλαδή σε μικρότερο αριθμό μεταβλητών

που ονομάζονται κύριες συνιστώσες, που είναι ασυσχέτιστες και μπορούν να ερμηνεύσουν το μεγαλύτερο ποσοστό διακυμάνσεων.

ΔΟΜΗ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Στο πρώτο μέρος της διπλωματικής αυτής θα ασχοληθούμε αναλυτικά με τις γενικές αρχές της Γραμμικής Άλγεβρας και τις εφαρμογές των αρχών αυτών στην ανάλυση δεδομένων. Στο δεύτερο μέρος θα ασχοληθούμε με το κομμάτι της ανάλυσης σε κύριες συνιστώσες. Και στο τρίτο μέρος με την εφαρμογή της Ανάλυσης Κυρίων Συνιστωσών στην αναγνώριση προσώπων. Στο τέταρτο μέρος δίνουμε ένα αναλυτικό παράδειγμα ανάλυσης κυρίων συνιστωσών.

Abstract

MULTIDIMENSIONAL DATA

The data we collect and need to process are by majority multidimensional. Multidimensional data analysis is particularly a difficult subject and the solution to this problem is achieved by dimension reduction techniques.

BENEFITS OF REDUCING DIMENSIONS

There are a variety of benefits from reducing data dimensions. Some of them are:

• Reducing data dimensions can lead us to more comprehensible models.

• Many data mining algorithms work best when there is a smaller number of dimensions.

• According to the "multiple dimensions curse", the larger the number of dimensions, the "thinner" the data becomes in the space they occupy, so we seek to reduce their dimensions.

DIMENSIONAL REDUCTION TECHNIQUES

Dimensional reduction techniques are categorized as linear and non-linear. By linear we mean the methods that are using Linear Algebra techniques such as vector spaces, linear functions, linear equation systems e.t.c. in order to project data from a multi-dimensional space into a space with a smaller number of dimensions. Such a linear dimensional reduction technique is what we called the **principal component analysis** (p.c.a)

PRINCIPAL COMPONENT ANALYSIS

It is one of the popular linear methods of data analysis for dimensional reduction. Principal Component Analysis is a statistical process that converts an initial group of observations of potentially correlated variables into a set of new observations with variables that are linear combinations of the initial ones, so as they are uncorrelated and contain as much variance of the initial variables as possible. Thus, we arrive at a smaller number of variables called main components, which are uncorrelated and can interpret the majority of the initial data variance.

STRUCTURE OF THESIS

In the first chapter of this thesis we will deal with the general principles of Linear Algebra and their applications in data analysis. In the second chapter we study the analysis in principal components. The third chapter deals, with the application of the Principal Component Analysis in Image Processing. In the fourth chapter we give a detailed example of principal components analysis.

Πίνακας περιεχομένων

Κεφάλαιο 1: Αρχές Γραμμικής Άλγεβρας

Στο κεφάλαιο αυτό δίνουμε μερικές από τις βασικές έννοιες της Γραμμικής Άλγεβρας τις οποίες θα χρησιμοποιήσουμε αργότερα.

1.1 Βασικές έννοιες (σώμα- διάνυσμα- διανυσματικός χώρος – υπόχωρος)

Σώμα είναι ένα σύνολο K στο οποίο ορίζονται δύο διμελής πράξεις, της πρόσθεσης και του πολλαπλασιασμού:

$$(\alpha,b) \rightarrow \alpha+b$$
$$(\alpha,b) \rightarrow \alpha \cdot b$$

που ικανοποιούν τα αξιώματα:

ΑΣ1. (προσεταιριστική ιδιότητα για την πρόσθεση και τον πολλαπλασιασμό) για κάθε $a, b, c \in K$: (a + b)+c = a + (b + c), (a· b)· c = a· (b · c)

ΑΣ2. **(**αντιμεταθετική ιδιότητα για την πρόσθεση και τον πολλαπλασιασμό): για κάθε a, b, c $\in K$, ισχύουν: a + b=b +a , a· b=b· a

ΑΣ3. **(**επιμεριστική ιδιότητα της πρόσθεσης ως προς τον πολλαπλασιασμό): για κάθε a, b, c $\in K$, ισχύει: a· (b + c)=a· b + a·c

ΑΣ4. Υπάρχουν στοιχεία 0 $\in K$ και 1$\in K$, τέτοια ώστε: για κάθε a $\in K$, a+0=a και a·1=a

ΑΣ5. Για κάθε a $\in K$ υπάρχει μοναδικό b $\in K$ τέτοιο ώστε a + b =0. Το μοναδικό στοιχείο b με αυτή την ιδιότητα συμβολίζεται -a και ονομάζεται αντίθετο του a.

ΑΣ6. Για κάθε a$\in K$, a≠0 υπάρχει μοναδικό b $\in K$ τέτοιο ώστε a • b =1. Το μοναδικό στοιχείο b με αυτή την ιδιότητα συμβολίζεται a^{-1} και ονομάζεται αντίστροφο του a.

Παράδειγμα 1.1.1

Οι ρητοί αριθμοί ¤ , οι πραγματικοί αριθμοί ¨ και οι μιγαδικοί αριθμοί £ αποτελούν σώματα (ικανοποιούν όλα τα παραπάνω).

Σε αντίθεση με τους ακέραιους αριθμούς Ά , όπου δεν ικανοποιείται το ΑΣ6, γιατί π.χ. ένα α=2, δεν υπάρχει ακέραιος b τέτοιος ώστε αb=1

Ορισμός 1.1.2 (Διανυσματικός χώρος)

Θεωρούμε ένα σώμα K . Ένα σύνολο V ονομάζεται **διανυσματικός χώρος** πάνω στο σώμα K αν:

- Το V είναι κλειστό ως προς την πρόσθεση δηλαδή αν u ,v∈ V, τότε u+ v ∈ V
- Ισχύει η αντιμεταθετική ιδιότητα για την πράξη της πρόσθεσης, δηλ.
  u+ v= v +u για όλα τα u, v ∈ V.
- Ισχύει η προσεταιριστική ιδιότητα για την πράξη της πρόσθεσης, δηλ.
  u +(v+ w) =(u +v) +w για όλα τα u, v, w ∈ V.
- Υπάρχει το ουδέτερο στοιχείο για την πράξη της πρόσθεσης, δηλ. υπάρχει
  0∈ V τέτοιο ώστε u+0=0+u=u για όλα τα u ∈ V.
- Για κάθε u ∈ V, υπάρχει στοιχείο (-u) ∈ V, τέτοιο ώστε u+(-u)=(-u)+ u =0.
- Για κάθε α ∈ K , u ∈ V ορίζεται ένα στοιχείο α·u ∈ V , το οποίο ονομάζεται βαθμωτό πολλαπλάσιο του u επί το α, και ισχύουν οι κάτωθι ιδιότητες:
  (α) α·(u+v)=α·u+α·v για όλα τα α∈ K , u,v∈V
  (β) (α +β) ·v=α·v+β·v για όλα τα α, β∈ K , v∈V
  (γ) α· (β·v)=(αβ) ·v για όλα τα α, β∈ K , v∈V
  (δ) 1·v=v για όλα τα v∈V .

Τα στοιχεία του συνόλου V λέγονται **διανύσματα**, ενώ εκείνα του K συντελεστές.

→ Με την πρόσθεση επιτυγχάνεται η απεικόνιση του V x V στον V.

→ Με το βαθμωτό πολλαπλασιασμό επιτυγχάνεται η απεικόνιση του K x V στον V.

Έστω $a_1, a_2, ..., a_n$ στοιχεία ενός συνόλου Ε. Το $\vec{u} = (a_1, a_2, ..., a_n)$ ονομάζεται **διάνυσμα** $n_{\text{όστις}}$ τάξης και τα στοιχεία $a_1, a_2, ..., a_n$ ονομάζονται **συντεταγμένες του διανύσματος.**

Τα **μοναδιαία διανύσματα** του χώρου $\mathbb{R}^3$ (τριών διαστάσεων) είναι

$$\vec{e}_1 = (1,0,0),\ \vec{e}_2 = (0,1,0),\ \vec{e}_3 = (0,0,1)$$

Πράξεις διανυσμάτων

πρόσθεση / άθροισμα δύο διανυσμάτων $\vec{x} = (x_1, x_2, ..., x_n)$ και $\vec{y} = (y_1, y_2, ..., y_n)$ είναι το διάνυσμα με αρχή το $\vec{x}$ και πέρας το $\vec{y}$, αν το $\vec{y}$ εφαρμοστεί στο πέρας του $\vec{x}$ και οι συντεταγμένες του είναι:

$$\vec{x} + \vec{y} = (x_1 + y_1,\ x_2 + y_2,\ ...,\ x_n + y_n)$$

αφαίρεση ονομάζεται η πρόσθεση ενός διανύσματος στο αντίθετο ενός άλλου διανύσματος. Το αντίθετο διάνυσμα συμβολίζεται με $-\vec{y}$ και έχουμε:

$$\vec{x} - \vec{y} = (x_1 - y_1,\ x_2 - y_2,\ ...,\ x_n - y_n)$$

πολλαπλασιασμός δύο διανυσμάτων είναι το λεγόμενο **εσωτερικό γινόμενο** δύο διανυσμάτων και ορίζεται να είναι ο αριθμός:

$$\vec{x} \cdot \vec{y} = (x_1 \cdot y_1,\ x_2 \cdot y_2,\ ...,\ x_n \cdot y_n)$$

γινόμενο διανύσματος επί ένα βαθμωτό k ($k \in$): έστω το διάνυσμα $\vec{x} = (x_1,\ x_2,\ ...,\ x_n)$ και το βαθμωτό $k \in$, τότε το γινόμενο:

$$k \cdot \vec{x} = (k \cdot x_1,\ k \cdot x_2,\ ...,\ k \cdot x_n)$$

Έστω τώρα ο διανυσματικός χώρος (διανυσμάτων) Ε πάνω στο σώμα K και $\{\vec{x}_1,\ \vec{x}_2, ..., \vec{x}_n\}$ ένα υποσύνολο του Ε .

Υπόχωρος του διανυσματικού χώρου Ε είναι κάθε μη κενό υποσύνολο του Ε για το οποίο ισχύουν οι παρακάτω προϋποθέσεις:

- το άθροισμα $\vec{x}_i + \vec{x}_j$ δύο διανυσμάτων του $\vec{x}_i$ και $\vec{x}_j$ περιέχεται στον υπόχωρο.
- Το γινόμενο $\alpha \vec{x}_i$ του διανύσματος του $\vec{x}_i$ επί το βαθμωτό $\alpha \in$ K περιέχεται στον υπόχωρο

Το σύνολο των γραμμικών συνδυασμών $x = a_1\vec{x}_1 + a_2\vec{x}_2 + ... + a_n\vec{x}_n$ με $a_i \in$K είναι λοιπόν ένας **υπόχωρος** V του Ε , δηλαδή:

$$V = \left\{ a_1\vec{x}_1 + a_2\vec{x}_2 + ... + a_n\vec{x}_n \ /\ \ a_i \in K \right\} ..$$

Τα διανύσματα $\vec{x}_1, \vec{x}_2, ..., \vec{x}_n$ του διανυσματικού χώρου Ε θα ονομάζονται **γραμμικά εξαρτημένα** όταν υπάρχουν βαθμωτά $k_1, k_2, ..., k_n$ όχι όλα 0 τέτοια ώστε:

$$k_1\vec{x}_1 + k_2\vec{x}_2 + ... + k_n\vec{x}_n = 0$$

Παράδειγμα 1.1.3

Τα παρακάτω διανύσματα είναι γραμμικά εξαρτημένα:

$\vec{x}_1$ =(1,2,3) $\vec{x}_2$ =(2,5,1) $\vec{x}_3$=(-1,-3,2)

Αρκεί να βρούμε τουλάχιστον ένα k_1, k_2, k_3 διάφορο του 0 τέτοιο ώστε:

$$k_1\vec{x}_1 + k_2\vec{x}_2 + k_3\vec{x}_3 = \vec{0} \Rightarrow$$

$$k_1(1,2,3) + k_2(2,5,1) + k_3(-1,-3,2) = \vec{0} \Rightarrow$$

$$\left.\begin{array}{l} k_1 + 2k_2 - k_3 = 0 \\ 2k_1 + 5k_2 - 3k_3 = 0 \\ 3k_1 + k_2 + 2k_3 = 0 \end{array}\right\} \Rightarrow \left.\begin{array}{l} k_1 = -2k_2 + k_3 \\ 2k_3 - 4k_2 + 3k_2 - 3k_3 = 0 \\ 3k_1 + k_2 + 2k_3 = 0 \end{array}\right\} \Rightarrow \left.\begin{array}{l} k_3 = k_2 \\ 3k_1 = 3k_2 \end{array}\right\} \Rightarrow k_1 = k_2$$

Επομένως, επειδή $k_1 = 1, k_2 = k_3 = -1$, τα διανύσματα είναι γραμμικά εξαρτημένα.

Παρατήρηση 1.1.4
Τα μοναδιαία διανύσματα του διανυσματικού χώρου $\mathbb{R}^3$ είναι γραμμικά ανεξάρτητα.

Βάση ενός διανυσματικού χώρου Ε είναι κάθε σύνολο διανυσμάτων $\{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}$ του που ικανοποιεί τις παρακάτω ιδιότητες:

- τα διανύσματα $\vec{x}_1, \vec{x}_2, ..., \vec{x}_n$ είναι γραμμικά ανεξάρτητα
- το τυχόν διάνυσμα $\vec{x}$ του χώρου Ε είναι δυνατόν να γραφεί σαν γραμμικός συνδυασμός των $\vec{x}_1, \vec{x}_2, ..., \vec{x}_n$ (το $\vec{x}$ λέμε ότι **παράγεται** από τα $\vec{x}_1, \vec{x}_2, ..., \vec{x}_n$) ή το σύνολο των διανυσμάτων του Ε είναι δυνατό**ν** να παραχθεί από το σύνολο $\{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}$.

Παράδειγμα 1.1.5
Τα μοναδιαία διανύσματα $\vec{e}_1, \vec{e}_2, \vec{e}_3$ αποτελούν βάση του χώρου $\mathbb{R}^3$ αφού:

- Είναι γραμμικά ανεξάρτητα

$$k_1\vec{e}_1 + k_2\vec{e}_2 + k_3\vec{e}_3 = \vec{0} \Rightarrow k_1(1,0,0) + k_2(0,1,0) + k_3(0,0,1) = (0,0,0) \Rightarrow k_1 = k_2 = k_3 = 0$$

- το κάθε διάνυσμα $\vec{x} = (x_1, x_2, x_3)$ του διανυσματικού χώρου $\mathbb{R}^3$ μπορεί να γραφεί κατά έναν και μοναδικό τρόπο ως γραμμικός συνδυασμός των μοναδιαίων διανυσμάτων οποιασδήποτε βάσης του.

$$\vec{x} = (x_1, x_2, x_3) = x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3$$

Ο αριθμός των στοιχείων της (οποιασδήποτε) βάσης ενός διανυσματικού χώρου Ε ονομάζεται **διάσταση του** συμβολίζεται με **dimE.**

1.2 Γραμμικές απεικονίσεις

Γραμμικές απεικονίσεις ονομάζονται οι συναρτήσεις από ένα διανυσματικό χώρο σε έναν άλλον οι οποίες διατηρούν αναλλοίωτη τη δομή του. Είναι οι απεικονίσεις που είναι συμβιβαστές με την **πρόσθεση** και το **βαθμωτό** πολλαπλασιασμό. Συγκεκριμένα η απεικόνιση f του διανυσματικού χώρου V στο διανυσματικό χώρο W (f: V → W) λέγεται **γραμμική** όταν:

$$f(\vec{x}+\vec{y}) = f(\vec{x}) + f(\vec{y}),\ \forall \vec{x}, \vec{y} \in V$$

$$f(\lambda \vec{x}) = \lambda f(\vec{x}) \quad \forall x \in V,\ \lambda \in K$$

Εάν f: E(dimE=n) → F(dimF=n) γραμμική απεικόνιση, τότε υπάρχει πίνακας $A^{n\times n}$ έτσι ώστε:

$$f(\vec{x}) = A\vec{x}$$

Π.χ. για τυχόν διάνυσμα $\vec{x}$ του χώρου $\mathbb{R}^3$, που είναι εφοδιασμένος με τη συνήθη βάση $\{\vec{e}_1, \vec{e}_2, \vec{e}_3\}$, θα έχουμε:

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3$$

Η εικόνα του $\vec{x}$ μέσω της f είναι το διάνυσμα $\vec{y}$:

$$\vec{y} = \begin{bmatrix} y_1 \\ \mathrm{M} \\ y_n \end{bmatrix} = f(\vec{x}) = f(x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3) = f(x_1\vec{e}_1) + f(x_2\vec{e}_2) + ... + f(x_n\vec{e}_n)$$

$$= x_1 f(\vec{e}_1) + x_2 f(\vec{e}_2) + x_3 f(\vec{e}_3)$$

Έστω ότι οι απεικονίσεις των μοναδιαίων διανυσμάτων του $\mathbb{R}^3$ μέσω της f είναι:

$$f(\vec{e}_1) = \begin{bmatrix} \alpha_{11} \\ \mathrm{M} \\ \alpha_{n1} \end{bmatrix} \quad f(\vec{e}_2) = \begin{bmatrix} a_{12} \\ \mathrm{M} \\ a_{n2} \end{bmatrix} \quad \dots \quad f(\vec{e}_n) = \begin{bmatrix} a_{1n} \\ \mathrm{M} \\ a_{nn} \end{bmatrix} \text{ τότε}$$

$$f(\vec{x}) = x_1 f(\vec{e}_1) + x_2 f(\vec{e}_2) + x_3 f(\vec{e}_3) =$$

$$= x_1 \begin{bmatrix} \alpha_{11} \\ \mathrm{M} \\ \alpha_{n1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ \mathrm{M} \\ a_{n2} \end{bmatrix} + \dots + x_n \begin{bmatrix} a_{1n} \\ \mathrm{M} \\ a_{nn} \end{bmatrix} =$$

$$= \begin{bmatrix} a_{11} & \text{L} & a_{1n} \\ \text{M} & \text{O} & \text{M} \\ a_{n1} & \text{L} & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \text{M} \\ x_n \end{bmatrix} = \text{A}\vec{x} = \text{L}_\text{A}\vec{x}$$

Παρατήρηση 1.2.1

Ο πίνακας Α εξαρτάται από τις βάσεις των διανυσματικών χώρων Ε και F και ορίζεται από τις εικόνες των διανυσμάτων της βάσης του χώρου Ε.

Έστω διανυσματικός χώρος Ε με dimE=n

Ταυτοτική απεικόνιση η ονομάζεται η απεικόνιση $I : E \xrightarrow{I} E$ που ορίζεται από την $\vec{x} \to I(\vec{x}) = \vec{x}$

Η ταυτοτική απεικόνιση είναι προφανώς γραμμική απεικόνιση με πίνακα που συμβολίζεται με Ι και ονομάζεται **μοναδιαίος** πίνακας (ο πίνακας που έχει στην κύρια διαγώνιο στοιχεία ίσα με 1 και αλλού τα στοιχεία είναι ίσα με 0).

1.3. Πράξεις πινάκων

Όπως είδαμε όταν στους διανυσματικούς χώρους άφιξης και αναχώρησης μιάς γραμμικής απεικόνισης έχουν οριστεί οι αντίστοιχες βάσεις, τότε στην απεικόνιση αντιστοιχεί ένας πίνακας.

(α) Πρόσθεση Αν στις γραμμικές απεικονίσεις f και g αντιστοιχούν οι ίδιων διαστάσεων πίνακες Α (p x n) και Β (p x n) τότε το άθροισμα τους f και g είναι και αυτό γραμμική απεικόνιση f+g του Ε στο F και σε αυτή αντιστοιχεί ο πίνακας Γ(p x n) =A (p x n) +B (p x n)

$$E \xrightarrow{f} E \Rightarrow A(p \times n) = a_{ij} \qquad i = 1,\text{K}\ , p \quad j = 1,\text{K}\ , n$$
$$E \xrightarrow{g} F \Rightarrow B(p \times n) = b_{ij} \qquad i = 1,\text{K}\ , p \quad j = 1,\text{K}\ , n$$

όπου dim(E)=n και dim(F)=p

$$E \xrightarrow{f+g} F \Rightarrow \Gamma\left(p \times n\right) = \gamma_{ij} = a_{ij} + b_{ij} \quad i = 1,\text{K}\ , p \quad j = 1,\text{K}\ , n$$

Παράδειγμα 1.3.1

Έστω ο 2x3 πίνακας $A=\begin{bmatrix} 1 & -2 & 3 \\ 2 & 1 & 4 \end{bmatrix}$ και ο 2x3 πίνακας $B=\begin{bmatrix} 2 & 3 & 1 \\ -1 & 2 & 1 \end{bmatrix}$

Το άθροισμα των πινάκων $A+B=\begin{bmatrix} 1+2 & -2+3 & 3+1 \\ -1+2 & 2+1 & 1+4 \end{bmatrix}=\begin{bmatrix} 3 & 1 & 4 \\ 1 & 3 & 5 \end{bmatrix}$

(β) Βαθμωτό γινόμενο πίνακα επί αριθμό

Αν στη γραμμική απεικόνιση f αντιστοιχεί ένα πίνακας Α (p x n) τότε το γινόμενο της f επί έναν αριθμό λ είναι μια γραμμική απεικόνιση λf του Ε στο F και σε αυτήν αντιστοιχεί ο πίνακας Γ(p x n) =λΑ (p x n)

$$E \xrightarrow{f} F \Rightarrow A(p\times n) = a_{ij} \quad i=1,\text{K}\ , p \quad j=1,\text{K}\ ,n$$
$$E \xrightarrow{\lambda f} F \Rightarrow \Gamma(p\times n) = \lambda \text{A} \Rightarrow \gamma_{ij} = \lambda a_{ij} \ \ i=1,\text{K}\ , p \quad j=1,\text{K}\ ,n$$

Παράδειγμα 1.3.2

Έστω ο 2x3 πίνακας Α=$\begin{bmatrix}1 & -2 & 3\\ 2 & 1 & 4\end{bmatrix}$ και η σταθερά λ=2

Ο πίνακας Γ= λ Α = 2 $\begin{bmatrix}1 & -2 & 3\\ 2 & 1 & 4\end{bmatrix}$ = $\begin{bmatrix}2 & -4 & 6\\ 4 & 2 & 8\end{bmatrix}$

(γ) Γινόμενο πινάκων

Αν στις γραμμικές απεικονίσεις f και g αντιστοιχούν οι πίνακες Α (p x n) και Β (q x p) τότε η σύνθεσή τους $g \text{ o } f$ είναι και αυτή γραμμική απεικόνιση του Ε στο G και σε αυτή αντιστοιχεί ο πίνακας $\Gamma(qxn) = B(qxp)xA(pxn)$

$$E \xrightarrow{f} F \Rightarrow A(p\times n) = a_{ij} \ \ i=1,\text{K}\ , p \quad j=1,\text{K}\ ,n$$
$$F \xrightarrow{g} G \Rightarrow B(n\times m) = b_{ij} \ \ i=1,\text{K}\ ,m \ \ j=1,\text{K}\ , p$$

$$E \xrightarrow{f} F \xrightarrow{g} G$$

$$\Gamma(qxn) = B(qxp)xA(pxn) \ = \Gamma_{ij} \quad i=1,\text{K}\ ,q \ , \ \ j=1,\text{K}\ ,n$$

$$\Gamma_{ij} = \sum_{k=1}^{p} \beta_{ik} a_{kj}$$

Παρατήρηση 1.3.3

Βασική προϋπόθεση για να υπολογιστεί το γινόμενο δύο πινάκων είναι: **το πλήθος των στηλών του πρώτου πίνακα να είναι ίσο με το πλήθος των γραμμών του δεύτερου πίνακα.** Οι γραμμές του νέου πίνακα θα είναι ίσες με το πλήθος των γραμμών του 1ου πίνακα και οι στήλες του θα είναι ίσες με το πλήθος των στηλών του 2ου πίνακα.

Παράδειγμα 1.3.4

Έστω ο πίνακας Α= $\begin{bmatrix}1 & 2 & -1\\ 0 & 2 & 1\end{bmatrix}$ και ο πίνακας Β= $\begin{bmatrix}1 & 2\\ -1 & 0\\ 2 & 1\end{bmatrix}$

Α (2x3) και ο Β (3x2)

Το πλήθος στηλών του Α είναι ίσο με το πλήθος των γραμμών του Β επομένως μπορούμε να πολλαπλασιάσουμε τους δύο πίνακα και ο νέος πίνακας θα έχει διαστάσεις: οι γραμμές του νέου πίνακα θα είναι ίσες με τις γραμμές του Α πίνακα (=2) και οι στήλες του νέου πίνακα ίσες με τις στήλες του πίνακα Β(=2)

$$AB=\begin{bmatrix} 1x1+2x(-1)+(-1)x2 & 1x2+2x0+(-1)x1 \\ 0x1+2x(-1)+1x2 & 0x2+2x0+1x1 \end{bmatrix} = \begin{bmatrix} -3 & 1 \\ 0 & 1 \end{bmatrix}$$

Το στοιχείο (1,1) αντιστοιχεί στο εσωτερικό γινόμενο της $1^{ης}$ γραμμής πίνακα Α και πρώτης στήλης πίνακα Β. Το στοιχείο (2,1) αντιστοιχεί στο εσωτερικό γινόμενο της $2^{ης}$ γραμμής πίνακα Α και πρώτης στήλης πίνακα Β κ.ο.κ.

1.4. Στοιχεία πινάκων

(α) Αντίστροφος πίνακα (A^{-1})

Αν Α είναι ένας τετραγωνικός πίνακας (n x n) , και εάν υπάρχει (ένας και μόνο) τετραγωνικός πίνακας (n x n) Β τέτοιος ώστε :

$$AB=BA=I_n$$

ο πίνακας Β λέγεται **αντίστροφός του Α** και συμβολίζεται με A^{-1} .

Για την εύρεση του αντιστρόφου μπορούμε ξεκινώντας από τον επαυξημένο πίνακα (Α|Ι) και ακολουθώντας γραμμοπράξεις να καταλήξουμε στον επαυξημένο ($I|A^{-1}$) που μας δίνει και τον αντίστροφο.

Παράδειγμα 1.4.1

Έστω ο τετραγωνικός πίνακας $A=\begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 0 \\ 1 & 3 & 1 \end{bmatrix}$. Για τον αντίστροφό του έχουμε:

βήμα 1ο: παίρνουμε τον πίνακα Α και δίπλα του γράφουμε τον αντίστοιχο μοναδιαίο

$$(A|I) = \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 3 & 1 & 0 & 0 & 1 \end{array}\right)$$

βήμα 2ο: εκτελούμε γραμμοπράξεις μέχρι στη θέση του πίνακα Α να εμφανιστεί ο μοναδιαίος //ο πίνακας που θα βρίσκεται στο τέλος της διαδικασίας, στη θέση που αρχικά ήταν ο μοναδιαίος θα είναι και ο αντίστροφος.

$$(A|I) = \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 3 & 1 & 0 & 0 & 1 \end{array}\right) \quad \gamma_2=\gamma_2-\gamma_1 \quad \gamma_3=\gamma_3-\gamma_1$$

$$(A|I) = \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 1 \end{array}\right) \quad \gamma_2 \leftrightarrow \gamma_3$$

$$(A|I) = \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 1 \\ 0 & -1 & -1 & -1 & 1 & 0 \end{array}\right) \quad \gamma_3 = \gamma_3 + \gamma_2$$

$$(A|I) = \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & -2 & 1 & 1 \end{array}\right) \quad \gamma_3 = (-1)\gamma_3$$

$$(A|I) = \left(\begin{array}{ccc|ccc} 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 2 & -1 & -1 \end{array}\right) \quad \gamma_1 = \gamma_1 - \gamma_3$$

$$(A|I) = \left(\begin{array}{ccc|ccc} 1 & 2 & 0 & -1 & 1 & 1 \\ 0 & 1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 2 & -1 & -1 \end{array}\right) \quad \gamma_1 = \gamma_1 - 2\gamma_2$$

$$(A|I) = \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 1 & -1 \\ 0 & 1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 2 & -1 & -1 \end{array}\right) = (I|A^{-1})$$

Άρα ο αντίστροφος είναι ίσος με $A^{-1} = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 0 & 1 \\ 2 & -1 & -1 \end{bmatrix}$

Παρατήρηση 1.4.2 (αντίστροφός διαγωνοποιημένου πίνακα)

Εάν $A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ (και $\lambda_1, \lambda_2 \neq 0$) τότε $A^{-1} = \begin{pmatrix} \frac{1}{\lambda_1} & 0 \\ 0 & \frac{1}{\lambda_2} \end{pmatrix}$

(β) Ανάστροφος πίνακα (A^T)

Εάν δίνεται ένας πίνακας $A = (a_{ij})$, τότε ο **ανάστροφός** του $A^T = (a_{ij}^T)$ είναι ο πίνακας που ορίζεται από τις σχέσεις $a_{ij}^T = a_{ji}$ Δηλαδή, για να κατασκευάσουμε τον πίνακα A^T θέτουμε τις **γραμμές του Α σαν στήλες του** A^T **και τις στήλες του Α σαν γραμμές του** A^T.

Παράδειγμα 1.4.3

Αν ο πίνακας $A = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 2 & 4 \end{bmatrix}$ είναι διάστασης 3x2

Ο ανάστροφος του $A^T=\begin{bmatrix} 1 & 0 \\ 2 & 2 \\ -1 & 4 \end{bmatrix}$ είναι διάστασης 2x3.

Σημείωση

Εάν $x=\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ τότε $x^T=\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$

(γ) Ίχνος τετραγωνικού πίνακα

Το άθροισμα των στοιχείων της κύριας διαγωνίου ενός τετραγωνικού πίνακα λέγεται **ίχνος/trace** του. Δηλαδή, εάν Α(nxn) τετραγωνικός, τότε

$$\textbf{ίχνος } A=\sum_{i=1}^{n} a_{ii}$$

Παράδειγμα 1.4.4

Αν $A=\begin{bmatrix} 2 & 3 & -1 \\ 2 & 4 & 2 \\ 6 & 3 & 1 \end{bmatrix}$ trace A= 2+4+1=7

(δ) Ορίζουσα πίνακα

Ορίζουσα ενός πίνακα Π_ν είναι μια απεικόνιση $D:\Pi_\nu \to K$ η οποία ικανοποιεί τις ιδιότητες:

- $D(A_1,...,cA_i,...,A_n) = c \quad D(A_1,...,A_i,...,A_n)$
- $D(I)=1$ όπου Ι μοναδιαίος

Η ορίζουσα ενός πίνακα Α συμβολίζεται με **|A| ή det A.**

Ορίζουσα 2x2

$$\det\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}\cdot a_{22}-a_{12}\cdot a_{21}$$

Ορίζουσα 3x3

$$\det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \alpha_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - \alpha_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + \alpha_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$= a_{11}(a_{22}\cdot a_{33}-a_{23}a_{32}) - a_{12}(a_{21}\cdot a_{33}-a_{23}a_{31}) + a_{13}(a_{21}\cdot a_{32}-a_{22}a_{31})$$

Παρατήρηση 1.4.5

(I) Η ορίζουσα ενός τριγωνικού (άνω ή κάτω) πίνακα ισούται με το γινόμενο των στοιχείων της κύριας διαγωνίου.

(II) Για να ελέγξουμε αν ένας πίνακας είναι **αντιστρέψιμος** αρκεί να δείξουμε ότι η ορίζουσα του είναι διάφορη του 0 , δηλαδή **det ≠ 0**

Παράδειγμα 1.4.6

Αντιστρέφεται ο πίνακας $A=\begin{bmatrix}2 & -1\\ 4 & -2\end{bmatrix}$;

Επειδή detA= 2x(-2)-4x(-1)= -4 +4= 0, ο πίνακας δεν είναι αντιστρέψιμος.

1.5. Ιδιοτιμές – ιδιοδιανύσματα πίνακα

Έστω η γραμμική απεικόνιση f του διανυσματικού χώρου E στον E με dim(E)=n.

$$E \xrightarrow{F} E$$

$$\vec{x} \to \vec{y} = f(\vec{x})$$

Το διάνυσμα $\vec{u}$ του E ονομάζεται **ιδιοδιάνυσμα** ή **χαρακτηριστικό διάνυσμα** της f και αντιστοιχεί στην **ιδιοτιμή** ή **χαρακτηριστική τιμή** $\lambda \in$ τότε και μόνο τότε:

$$f(\vec{u})=\lambda\vec{u}$$

Αν ο A είναι ο πίνακας που αντιστοιχεί στη γραμμική απεικόνιση f για μια επιλεγμένη βάση του E θα έχουμε:

$$f(\vec{u})=\lambda\vec{u} \quad \to \quad A\vec{u}=\lambda\vec{u} \to \quad A\vec{u}-\lambda\vec{u}=0 \to \quad (A-\lambda)\vec{u}=0 \to |A-\lambda I|=0$$

Οι τιμές του λ που επαληθεύουν την $|A-\lambda I|=0$ είναι οι **ιδιοτιμές** του A και σε κάθε μια από αυτές αντιστοιχεί ένα **ιδιοδιάνυσμα.**

Πλήθος ιδιοτιμών τετραγωνικού πίνακα

Σε κάθε τετραγωνικό πίνακα A (nxn) αντιστοιχούν το πολύ n πραγματικές ιδιοτιμές που είναι οι ρίζες της εξίσωσης $|A-\lambda I|=0$, δηλαδή οι ιδιοτιμές προκύπτουν από την επίλυση της εξίσωσης:

$$\Delta=\det\begin{pmatrix} a_{11}-\lambda & a_{12} & \text{L} & a_{1n} \\ a_{21} & a_{22}-\lambda & \text{L} & a_{2n} \\ \text{M} & \text{L} & \text{O} & \text{M} \\ a_{n1} & a_{n2} & \text{L} & a_{nn}-\lambda \end{pmatrix}=0$$

Παρατήρηση 1.5.1

Εάν ο Α είναι **συμμετρικός** πίνακας (δηλαδή τα συμμετρικά, ως προς την κύρια διαγώνιο του στοιχεία, είναι ίσα, τότε:

- όλες οι ρίζες της Δ=0 είναι πραγματικές, δηλαδή ο Α έχει n πραγματικές χαρακτηριστικές τιμές.
- Σε κάθε απλή χαρακτηριστική τιμή λ αντιστοιχεί ένας χαρακτηριστικός υπόχωρος μια διάστασης.
- Σε κάθε πολλαπλή ρίζα λ της detA=0, πολλαπλή χαρακτηριστική τιμή του Α αντιστοιχεί ένας χαρακτηριστικός υπόχωρος με διάσταση όση η πολλαπλότητα της ρίζας λ.
- Τα χαρακτηριστικά διανύσματα που αντιστοιχούν σε άνισες χαρακτηριστικές τιμές είναι γραμμικά ανεξάρτητα.

Σύμφωνα με την τελευταία ιδιότητα μπορούμε να θεωρούμε σα βάση διανυσματικού χώρου Ε την $\{u_i\}$, όπου

$$A\vec{u}_i = \lambda_i \vec{u}_i \quad \text{i=1,2,...,n}$$

όπου λ_i οι διαφορετικές ιδιοτιμές του πίνακα Α.

Παράδειγμα 1.5.2

Να υπολογιστούν οι ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα:

$$\text{A}=\begin{bmatrix} 2 & 1 & 1 \\ 2 & 3 & 2 \\ 3 & 3 & 4 \end{bmatrix}$$

Οι ιδιοτιμές δίνονται από την επίλυση της εξίσωσης det(A-λI)=0 $\Rightarrow$

$$\det\begin{bmatrix} 2-\lambda & 1 & 1 \\ 2 & 3-\lambda & 2 \\ 3 & 3 & 4-\lambda \end{bmatrix}=0 \Rightarrow$$

$$(2-\lambda)\det\begin{bmatrix} 3-\lambda & 2 \\ 3 & 4-\lambda \end{bmatrix} -1\det\begin{bmatrix} 2 & 2 \\ 3 & 4-\lambda \end{bmatrix} + 1\det\begin{bmatrix} 2 & 3-\lambda \\ 3 & 3 \end{bmatrix}=0 \Rightarrow$$

$(2-\lambda)((3-\lambda)(4-\lambda)-6)-(2(4-\lambda)-6)+(6-3(3-\lambda))=0 \Rightarrow$

$(2-\lambda)(12-3\lambda-4\lambda+\lambda_2-6)-(8-2\lambda-6)+(6-9+3\lambda)=0 \Rightarrow -\lambda^3+9\lambda^2-15\lambda+7=0$

Από τη μέθοδο/σχήμα Horner βρίσκουμε ότι οι τρεις ιδιοτιμές είναι λ=1(διπλή) και λ=7.

Για **λ=7**

$$\begin{bmatrix} -5 & 1 & 1 \\ 2 & -4 & 2 \\ 3 & 3 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow$$

$$-5x_1 + x_2 + x_3 = 0 \Rightarrow x_3 = 5x_1 - x_2$$

$$2x_1 - 4x_2 + 2x_3 = 0 \Rightarrow 2x_1 - 4x_2 + 2(5x_1 - x_2) = 0 \Rightarrow 2x_1 - 4x_2 + 10x_1 - 2x_2 = 0 \Rightarrow x_2 = 2x_1$$

$$3x_1 + 3x_2 - 3x_3 = 0 \Rightarrow x_3 = 5x_1 - 2x_1 = 3x_1$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ 2x_1 \\ 3x_1 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Δηλαδή στην ιδιοτιμή λ=7 αντιστοιχεί το ιδιοδιάνυσμα $\vec{u}$=(1,2,3)

Για **λ=1**

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \quad x_1 + x_2 + x_3 = 0 \Rightarrow x_1 = -x_2 - x_3$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -x_2 - x_3 \\ x_2 \\ x_3 \end{bmatrix} = x_2 \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

Στην ιδιοτιμή 1 αντιστοιχούν δύο ιδιοδιανύσματα :

$$\vec{u}_1 = (-1,1,0) \text{ και } \vec{u}_2 = (-1,0,1)$$

Βλέπουμε ότι ο υπόχωρος, που αντιστοιχεί στην ιδιοτιμή 1, έχει διάσταση όση και η πολλαπλότητα της ρίζας.

1.6. Διαγωνοποίηση τετραγωνικού πίνακα

Έστω η γραμμική απεικόνιση f του διανυσματικού χώρου Ε στον Ε με dim(E)=n, δηλαδή $E \xrightarrow{f} E$ με $\vec{x} \to \vec{y}$= f($\vec{x}$). Αν Α ο πίνακας που αντιστοιχεί στην f και λ_i οι k ιδιοτιμές του τιμές τότε ο διανυσματικός χώρος του Ε μπορεί να θεωρηθεί σαν άθροισμα των χαρακτηριστικών υποχώρων $E_i \quad i=1,2,...,k$ (υπόχωροι που παράγονται από τα ιδιοδιανύσματα του Α)

$$E = \oplus\{E_i / i = 1,2,...,k\}$$

Μπορούμε, όπως αναφέραμε, να θεωρούμε σα βάση διανυσματικού χώρου Ε την $\{u_i\}$, όπου $A\vec{u}_i = \lambda_i \vec{u}_i$ i=1,2,...,n και όπου λ_i οι διαφορετικές ιδιοτιμές του πίνακα Α.

Ο πίνακας, τότε, που αντιστοιχεί στην γραμμική απεικόνιση f μπορεί να πάρει τη μορφή:

$$\Delta = \begin{pmatrix} \lambda_1 & & 0 \\ & O & \\ 0 & & \lambda_n \end{pmatrix}$$

Είναι ουσιαστικά ένας διαγώνιος πίνακας όπου στην κύρια διαγώνιο του έχει σαν τιμές τις ιδιοτιμές του Α. Τότε

$$A = U^{-1}\Delta U$$

όπου η *i*-οστή στήλη του U^{-1} είναι οι συντεταγμένες του ιδιοδιανύσματος $\vec{u_i}$ ως προς την αρχική βάση του Ε (δηλαδή ο U ο πίνακας αλλαγής βάσης). Η παραπάνω μορφή ονομάζεται **διαγωνοποίηση** του Α.

Παράδειγμα 1.6.1

Να διαγωνοποιηθεί ο παρακάτω πίνακας:

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 2 & 3 & 2 \\ 3 & 3 & 4 \end{bmatrix}$$

Σε προηγούμενο παράδειγμα βρήκαμε ότι οι ιδιοτιμές είναι λ=1 (διπλή), λ=7 και τα ιδιοδιανύσματα: (α) για λ=7 είναι $\vec{u}_1$ =(1,2,3) (β) για λ=1 (διπλή) είναι $\vec{u}_2$ =(1,0,-1) και $\vec{u}_3$ =(0,1,-1).

Ο πίνακας U (αλλαγής βάσης) είναι:

$$U=\begin{bmatrix}1 & 1 & 0\\ 2 & 0 & 1\\ 3 & -1 & -1\end{bmatrix}$$

και επειδή det U=6≠0 μπορούμε να τον αντιστρέψουμε (ο αντίστροφός του υπολογίζεται στη συνέχεια με τη βοήθεια γραμμοπράξεων)

$$\left(\begin{array}{ccc|ccc}1 & 1 & 0 & 1 & 0 & 0\\ 2 & 0 & 1 & 0 & 1 & 0\\ 3 & -1 & -1 & 0 & 0 & 1\end{array}\right) \qquad \gamma_2=\gamma_2-2\gamma_1 \quad \text{και} \quad \gamma_3=\gamma_3-3\gamma_1$$

$$\left(\begin{array}{ccc|ccc}1 & 1 & 0 & 1 & 0 & 0\\ 0 & -2 & 1 & -2 & 1 & 0\\ 0 & -4 & -1 & -3 & 0 & 1\end{array}\right) \qquad \gamma_3=\gamma_3-2\gamma_2$$

$$\left(\begin{array}{ccc|ccc}1 & 1 & 0 & 1 & 0 & 0\\ 0 & -2 & 1 & -2 & 1 & 0\\ 0 & 0 & -3 & 1 & -2 & 1\end{array}\right) \qquad \gamma_2=\gamma_2+\frac{1}{3}\gamma_3$$

$$\left(\begin{array}{ccc|ccc}1 & 1 & 0 & 1 & 0 & 0\\ 0 & -2 & 0 & \frac{-5}{3} & \frac{1}{3} & \frac{1}{3}\\ 0 & 0 & -3 & 1 & -2 & 1\end{array}\right) \qquad \gamma_1=\gamma_1+\frac{1}{2}\gamma_2$$

$$\left(\begin{array}{ccc|ccc}1 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6}\\ 0 & -2 & 0 & \frac{-5}{3} & \frac{1}{3} & \frac{1}{3}\\ 0 & 0 & -3 & 1 & -2 & 1\end{array}\right) \qquad \gamma_2=\frac{-1}{2}\gamma_2 \quad \text{και} \quad \gamma_3=\frac{-1}{3}\gamma_3$$

$$\left(\begin{array}{ccc|ccc}1 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6}\\ 0 & 1 & 0 & \frac{5}{6} & \frac{-1}{6} & \frac{-1}{6}\\ 0 & 0 & 1 & \frac{-1}{3} & \frac{-2}{3} & \frac{-1}{3}\end{array}\right)$$

Είναι λοιπόν $$U^{-1}=\frac{1}{6}\begin{bmatrix}1 & 1 & 1\\ 5 & -1 & -1\\ -2 & 4 & -2\end{bmatrix}$$

και ο διαγωνοποιημένος πίνακας Λ του Α είναι

$$\Lambda=U^{-1}AU=\frac{1}{6}\begin{bmatrix}1 & 1 & 1\\ 5 & -1 & -1\\ -2 & 4 & -2\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 7\end{bmatrix}\begin{bmatrix}1 & 1 & 0\\ 2 & 0 & 1\\ 3 & -1 & -1\end{bmatrix}$$

1.7 Συμμετρικός πίνακας, ίχνος πίνακα – διακύμανση , συνδιακύμανση.

Συμμετρικός ονομάζεται ένας τετραγωνικός πίνακας $A(n,n)=\alpha_{ij}$ αν $a_{ij}=a_{ji}$, $i,j=1,2,....,n$ δηλαδή όταν τα συμμετρικά, ως προς την κύρια διαγώνιο του στοιχεία, είναι ίσα.

Ο πίνακας που προκύπτει από το γινόμενο ενός πίνακα (δεξιά ή αριστερά) με τον ανάστροφο του είναι συμμετρικός αλλά το γινόμενο των πινάκων AA^T διαφέρει από τον A^TA.

Παράδειγμα 1.7.1

Δίνεται ο πίνακας

$$A=\begin{bmatrix}1 & 2\\ 3 & 5\\ 4 & 1\end{bmatrix} \qquad \text{Ο ανάστροφος } A^T=\begin{bmatrix}1 & 3 & 4\\ 2 & 5 & 1\end{bmatrix}$$

$$AA^T=\begin{bmatrix}1\times1+2\times2 & 1\times3+2\times5 & 1\times4+2\times1\\ 3\times1+5\times2 & 3\times3+5\times5 & 3\times4+5\times1\\ 4\times1+1\times2 & 4\times3+1\times5 & 4\times4+1\times1\end{bmatrix}=\begin{bmatrix}5 & 13 & 6\\ 13 & 34 & 17\\ 6 & 17 & 17\end{bmatrix}$$

Βλέπουμε ότι: $a_{12}=a_{21},\ a_{13}=a_{31},\ a_{23}=a_{32}$ άρα ο AA^T είναι συμμετρικός. Ακόμα

$$A^TA=\begin{bmatrix}1\times1+3\times3+4\times4 & 1\times2+3\times5+4\times1\\ 2\times1+5\times3+1\times4 & 2\times2+5\times5+1\times1\end{bmatrix}=\begin{bmatrix}26 & 21\\ 21 & 30\end{bmatrix}$$

Δηλαδή και ο πίνακας A^TA είναι συμμετρικός.. Παρατηρούμε ότι οι δύο αυτοί πίνακες διαφέρουν.

Πρόταση 1.7.2

(α) **Ίχνος** AA^T **= Ίχνος** A^TA

(β) Το γινόμενο ενός πίνακα Α με τον ανάστροφο του είναι τετραγωνικός συμμετρικός πίνακας με ίχνος το άθροισμα των τετραγώνων όλων των στοιχείων του Α.

Έστω ο πίνακας

$$A=\begin{bmatrix}\alpha_1 & \alpha_2 & \alpha_3\\ b_1 & b_2 & b_3\\ c_1 & c_2 & c_3\\ d_1 & d_2 & d_3\end{bmatrix} \quad \text{οπότε} \quad A^T=\begin{bmatrix}a_1 & b_1 & c_1 & d_1\\ a_2 & b_2 & c_2 & d_2\\ a_3 & b_2 & c_3 & d_3\end{bmatrix}$$

AA^T =

$$\begin{bmatrix} \alpha_1^2+\alpha_2^2+\alpha_3^2 & \alpha_1 b_1+\alpha_2 b_2+a_3 b_3 & \alpha_1 c_1+\alpha_2 c_2+a_3 c_3 & \alpha_1 d_1+\alpha_2 d_2+a_3 d_3 \\ \alpha_1 b_1+\alpha_2 b_2+a_3 b_3 & b_1^2+b_2^2+b_3^2 & b_1 c_1+b_2 c_2+b_3 c_3 & b_1 d_1+b_2 d_2+b_3 d_3 \\ \alpha_1 c_1+\alpha_2 c_2+a_3 c_3 & b_1 c_1+b_2 c_2+b_3 c_3 & c_1^2+c_2^2+c_3^2 & c_1 d_1+c_2 d_2+c_3 d_3 \\ \alpha_1 d_1+\alpha_2 d_2+a_3 d_3 & b_1 d_1+b_2 d_2+b_3 d_3 & c_1 d_1+c_2 d_2+c_3 d_3 & d_1^2+d_2^2+d_3^2 \end{bmatrix}$$

Η **διακύμανση,** όπως ξέρουμε, είναι άθροισμα τετραγώνων, επομένως τα στοιχεία της κύριας διαγωνίου ενός τετραγωνικού συμμετρικού πίνακα ονομάζονται **στοιχεία της διακύμανσης του.** Τα υπόλοιπα στοιχεία του πίνακα που είναι συμμετρικά ως προς αυτήν ονομάζονται **στοιχεία της συνδιακύμανσης** του.

1.8 Συσχέτιση Γραμμικής Άλγεβρας με Ανάλυση Κύριων Συνιστωσών.

Έστω ότι δίνεται ένας πίνακας δεδομένων A.

Πίνακες διακύμανσης ή συνδιακύμανσης ή πίνακες αδράνειας ονομάζονται οι πίνακες AA^T **,** $A^T A$.

Οι πίνακες AA^T και $A^T A$ αν και συνήθως διαφέρουν προσδιορίζουν την ίδια κατάσταση (τη λεγόμενη ελλειψοειδές της αδράνειας του Α) του προβλήματος που αναφέρεται ο Α.

Διαγωνοποίηση πίνακα – ιδιοτιμές, ιδιοδιανύσματα πίνακα

Η **αδράνεια** του πίνακα που περιγράφει ένα φαινόμενο/σύστημα παίζει σημαντικό ρόλο στην ανάλυση του, με μεθόδους ανάλυσης δεδομένων.

Με τη διαγωνοποίηση ενός τετραγωνικού συμμετρικού πίνακα επιτυγχάνεται ο προσδιορισμός των **αξόνων συμμετρίας του ελλειψοειδούς της αδράνειας** που είναι οι ονομαζόμενοι παραγοντικοί άξονες. Υπολογίζονται επίσης και οι αντίστοιχες ιδιοτιμές $\lambda_1, \lambda_2, ..., \lambda_p$ του πίνακα.

Τη διαγωνοποίηση ενός πίνακα μπορούμε ακόμα να τη θεωρήσουμε και σαν μια πράξη που μηδενίζει την συνδιακύμανση και αφήνει αμετάβλητη τη διακύμανση. Τα δεδομένα που μας δίνονται μεταβάλλονται γύρω από τη μέση τους τιμή. Στην ανάλυση δεδομένων το μέσο σημείο Ο, σε ένα χώρο n διαστάσεων, δεν είναι η αρχή των αξόνων αλλά το μέσο σημείο όλων των παρατηρήσεων.

Κάθε παρατήρηση παρουσιάζει μια απόκλιση από αυτό το σημείο. Επομένως το σύνολο των παρατηρήσεων θα ορίζει ένα νέφος σημείων με κέντρο το μέσο σημείο Ο.

Η διεύθυνση ως προς την οποία αποκλίνει περισσότερο το νέφος από το σημείο Ο, προσδιορίζει την επικρατέστερη τάση απομάκρυνσης από τη μέση κατάσταση των παρατηρήσεων. Αυτή η διεύθυνση προσδιορίζει τον μεγαλύτερο άξονα του ελλειψοειδούς.

Οι άλλοι άξονες λαμβάνονται κατά φθίνουσα σειρά μεγέθους και δίνουν άλλες τάσεις λιγότερο έντονες από την πρώτη.

Αυτούς τους άξονες τους προσδιορίζουμε, όπως θα δούμε στη συνέχεια, διαγωνοποιώντας τον τετραγωνικό συμμετρικό πίνακα $A^T A$ (ή AA^T).

Κεφάλαιο 2: Ανάλυση Κύριων Συνιστωσών (Α.Κ.Σ.)
(Principal Component Analysis - PCA)

2.1 Εισαγωγή

Όταν συλλέγουμε πολυμεταβλητά δεδομένα - δηλαδή εξετάζουμε ένα σύνολο περιπτώσεων/αντικειμένων (cases) ως προς ένα σύνολο **ποσοτικών** μεταβλητών του ενδιαφέροντός μας - μερικές από τις μεταβλητές ίσως συσχετίζονται με αποτέλεσμα να έχουμε πλεονασμό (περιττότητα) της πληροφορίας που παρέχουν. Σκοπός της λεγόμενης **Ανάλυσης Κύριων Συνιστωσών** (Α.Κ.Σ.) είναι η μείωση του όγκου των δεδομένων χωρίς να έχουμε σημαντική απώλεια της πληροφορίας.

Τα δεδομένα μας, σ' ένα πρόβλημα Α.Κ.Σ., μπορούμε να θεωρήσουμε ότι σχηματίζουν ένα πίνακα με οριζόντιες γραμμές τις τιμές κάθε μιάς των **n** περιπτώσεών/αντικειμένων **(cases)** μας ως προς (κάθε μια από) τις μεταβλητές και στήλες τις **p** μεταβλητές του ενδιαφέροντός μας. Ο πίνακας των δεδομένων μας δηλαδή έχει την παρακάτω γενική μορφή.

		Μεταβλητές (variables)			
		M_1	M_2	…	M_p
Περιπτώσεις (cases)	x_1	α_{11}	a_{12}	…	α_{1p}
	x_2	α_{21}	α_{22}	…	α_{2p}
	…	…	…	…	…
	x_n	α_{n1}	α_{n2}	…	α_{np}

Μπορούμε λοιπόν να θεωρήσουμε ότι το i-άτομο (στην πραγματικότητα οι μετρήσεις που αφορούν αυτό το άτομο και παριστάνονται με τη βοήθεια του διανύσματος $\left(a_{i1}, a_{i2}, ..., a_{ip}\right)$), είναι ένα σημείο στον p – διάστατο χώρο (p - ο αριθμός των μεταβλητών). Οι παρατηρήσεις αποτελούν το λεγόμενο «**νέφος**» των σημείων (ανάλογα ορίζεται και το νέφος των μεταβλητών).

Η Α.Κ.Σ. είναι μια πολυμεταβλητή (στατιστική) ανάλυση αυτού του πίνακα. Μετασχηματίζουμε τον πίνακα αυτό σ' έναν νέο πίνακα με ορθογώνιες στήλες

(μεταβλητές) οι οποίες ονομάζονται **κύριες συνιστώσες (principal components) ή κύριοι παράγοντες (factors).** Με άλλα λόγια, η Α.Κ.Σ. προσπαθεί να ανακαλύψει (διαισθητικά) τη δομή του νέφους (ομοιότητα των παρατηρήσεων) των σημείων προβάλλοντάς το σε υπόχωρο μικροτέρων διαστάσεων (συνήθως 2 έως 5) με ορθογώνιους άξονες **(παραγοντικοί άξονες).** Ουσιαστικά προσδιορίζει τη διεύθυνση αυτών των αξόνων και καθορίζει τις νέες συντεταγμένες των σημείων του νέφους αναφορικά με τους νέους άξονες.

Είναι η παλαιότερη, απλή και πιο δημοφιλής Στατιστική τεχνική. Ξεκινά από τον Cauchy, αλλά θεμελιώθηκε από τον Pearson. Γενίκευσής της είναι η Ανάλυση Αντιστοιχιών (ποιοτικά δεδομένα). Εφαρμογές της μεθόδου μπορεί να βρει κανείς στις νευροεπιστήμες, υπολογιστές, βιολογία, οικονομία, οικολογία, Αρχιτεκτονικά κ.λ.π.

2.2 Πλεονεκτήματα της μεθόδου

Με την Α.Κ.Σ. δημιουργούμε ένα σύνολο γραμμικών συνδυασμών (νέες μεταβλητές) των αρχικών συσχετισμένων μεταβλητών, έτσι ώστε οι συνδυασμοί να είναι ασυσχέτιστες μεταβλητές και να ερμηνεύουν (περιέχουν) όσο γίνεται μεγαλύτερο μέρος της διακύμανσης των (αρχικών) δεδομένων μας. Αυτό έχει σαν πλεονέκτημα τα παρακάτω.

- Πολλές φορές η αλληλεξάρτηση των μεταβλητών παίζει σημαντικότατο ρόλο στη στατιστική ανάλυση των δεδομένων μας. Δημιουργούμε λοιπόν ένα σύνολο ασυσχέτιστων μεταβλητών, απαραίτητο σε αρκετές στατιστικές μεθόδους (π.χ. στην πολλαπλή παλινδρόμηση υπάρχει το πρόβλημα της πολυσυγγραμμικότητας – συσχέτισης των ανεξαρτήτων μεταβλητών — με αποτέλεσμα μεροληψία στις εκτιμήσεις της εξαρτημένης μεταβλητής).
- Αν με τις κύριες συνιστώσες ερμηνεύουμε μεγάλο μέρος της διακύμανσης των δεδομένων μας, τότε έχουμε πετύχει περιορισμό των διαστάσεων του προβλήματός μας. Οι συνιστώσες αυτές είναι οι διευθύνσεις στις οποίες υπάρχει η περισσότερη διασπορά των δεδομένων, δηλαδή αυτές που τα σημεία είναι πιο διασπαρμένα. Εάν λοιπόν ήθελε να αποθηκεύσει κανείς τα αρχικά δεδομένα, θα ήταν απαραίτητη μια τεράστια ίσως βάση δεδομένων. Με την εφαρμογή της Α.Κ.Σ. αποθηκεύουμε λιγότερες μεταβλητές με

αποτέλεσμα να χρειάζεται μικρότερος χώρος αποθήκευσης και να αυξάνεται η ταχύτητα επεξεργασίας των δεδομένων.

- Πολλές φορές, έχουμε λίγες περιπτώσεις προς εξέταση, αλλά πολλές μεταβλητές (π.χ ασθενείς στην Ιατρική, αρχαιολογικά ευρήματα στην Αρχαιολογία). Στις περιπτώσεις αυτές η μείωση του αριθμού των διαστάσεων του προβλήματος είναι απαραίτητη για την εξαγωγή χρήσιμων συμπερασμάτων. Είναι η μόνη λύση για να μπορεί να κάνει κανείς κάποιου είδους Στατιστική ανάλυση στα δεομένα.
- Δημιουργούμε όπως αναφέραμε νέες (ασυσχέτιστες) μεταβλητές σαν γραμμικούς συνδυασμούς των παλαιότερων μεταβλητών. Στις (νέες αυτές) μεταβλητές δίνουμε νέα ονόματα, σχετίζοντας τις νέες μεταβλητές με τις παλιές. Ο συσχετισμός αυτός είναι απαραίτητος για την ερμηνεία των νέων μεταβλητών (χρήσιμο στην Ψυχιατρική, Ψυχολογία κλπ).
- Με την Α.Κ.Σ. μπορούμε να ανακαλύψουμε ομοειδείς ομάδες παρατηρήσεων, οι οποίες δεν ήταν γνωστές από πριν. Ακόμα μπορούμε να παρατηρήσουμε τάσεις ανάμεσα στα δεδομένα, outliers κ.λ.π. Μπορούμε να βρούμε σχέσεις μεταξύ των παρατηρήσεων ή ακόμα και των μεταβλητών όχι φανερές από πριν.

Η ΑΚΣ, από άποψη:

Στατιστική: είναι ο υπολογισμός ευθειών, επιπέδων ή υπερεπιπέδων στον p–διάστατο χώρο που εκτιμούν (προσεγγίζουν) όσο το δυνατόν καλύτερα (υπό την έννοια των ελαχίστων τετραγώνων) τα δεδομένα. Είναι γνωστό ότι η ευθεία ή το επίπεδο αυτό (εκτιμητής ελαχίστων τετραγώνων) κάνει τη διασπορά των σημείων στην ευθεία ή στο επίπεδο όσο το δυνατόν μεγαλύτερη. Μια κύρια συνιστώσα δεν είναι αρκετή για να περιγράψει/μοντελοποιήσει τη διαφοροποίηση των δεδομένων.

Γραμμικής άλγεβρας: θέλουμε να υπολογίσουμε την πιο «σημαντική» βάση για να εκφράσουμε ξανά ένα παραποιημένο (από το θόρυβο) σύνολο δεδομένων. Με τη βοήθειά της ανακαλύπτουμε τις «κρυμμένες» δυναμικές ενός συστήματος σημείων.

2.3. Προϋποθέσεις εφαρμογής της μεθόδου

Η Α.Κ.Σ. εφαρμόζεται χωρίς καμία εκ των προτέρων υπόθεση, για το ποιές μεταβλητές ή ποιά αποτελέσματα παίζουν σημαντικότερο ρόλο στο φαινόμενο που περιγράφει ο πίνακας που αναλύεται. Μερικές προϋποθέσεις για την εφαρμογή της είναι και οι:

- Το μέγεθος δείγματος (περιπτώσεων/αντικειμένων) πρέπει να είναι τουλάχιστον 50 (αν και μερικοί συγγραφείς προτιμούν μέγεθος τουλάχιστον 100).
- Οι περιπτώσεις πρέπει να είναι ανεξάρτητες μεταξύ τους. Αν υπάρχουν outliers, θα πρέπει να απομακρύνονται.
- Οι μεταβλητές θα πρέπει να είναι ποσοτικές, όχι απαραίτητα με την ίδια μονάδα μέτρησης. Αν είναι κανονικοποιημένες, η λύση τους προβλήματος ισχυροποιείται.

2.4 Αναφορικά με τη μέθοδο της Α.Κ.Σ.

Α. Ιδέα της μεθόδου

Εάν A είναι ένας τετραγωνικός συμμετρικός πίνακας διαστάσεων pxp, τότε ο A μπορεί να γραφεί σα γινόμενο:

$$A = P\Delta P^T$$

όπου Δ ένας διαγώνιος πίνακας pxp με στοιχεία της διαγωνίου τις ιδιοτιμές του πίνακα A, P ένας ορθογώνιος πίνακας pxp, με στήλες τα κανονικοποιημένα ιδιοδιανύσματα των αντίστοιχων ιδιοτιμών του A (φασματική ανάλυση του πίνακα).

Με τη βοήθεια της παραπάνω σχέσης έχουμε:

$$A = P\Delta P^T \Rightarrow P^{-1}AP = P^{-1}P\Delta P^T P \overset{P^{-1}=P^T}{\Rightarrow} \Delta = P^{-1}AP$$

δηλαδή

$$\Delta = P^{-1}AP$$

Η παραπάνω διαδικασία ονομάζεται **διαγωνοποίηση** του πίνακα A (ξεκινώντας από έναν πίνακα A καταλήγουμε σ' έναν διαγώνιο πίνακα Δ).

Στην περίπτωση της Α.Κ.Σ. έχουμε έναν πίνακα δεδομένων X, ο δε **πίνακας συσχετίσεων** των δεδομένων/μεταβλητών είναι ο πίνακας $X^T X$. Θα θέλαμε να υπολογίσουμε έναν πίνακα U, ο οποίος πολλαπλασιαζόμενος με τον πίνακα X τον μετασχηματίζει σ' έναν νέο πίνακα $Y = XU$ δεδομένων/μεταβλητών τα οποία είναι ασυσχέτιστα, δηλαδή ο πίνακας συσχετίσεων των νέων μεταβλητών είναι διαγώνιος. Τη διαδικασία υπολογισμού του πίνακα U τη δίνει η διαγωνοποίηση του πίνακα $X^T X$. Πράγματι, έστω ότι ο πίνακας $X^T X$ διαγωνοποιείται με τη μορφή:

$$X^T X = U \Delta U^T$$

όπου Δ ένας διαγώνιος πίνακας pxp με στοιχεία της διαγωνίου τις ιδιοτιμές του πίνακα $X^T X$, P ένας ορθογώνιος πίνακας pxp, με στήλες τα κανονικοποιημένα ιδιοδιανύσματα των αντίστοιχων ιδιοτιμών (του πίνακα $X^T X$). Τότε ο πίνακας ο πίνακας συσχετίσεων των νέων μεταβλητών Y είναι ο πίνακας:

$$Y^T Y = \left(XU\right)^T XU = U^T X^T XU = U^T \left(U \Delta U^T\right) U = \Delta$$

δηλαδή διαγώνιος (σημαίνει ότι οι μεταβλητές είναι ανά δύο ασυσχέτιστες).

Παράδειγμα 2.4.1
Δίνεται ο παρακάτω πίνακας δεδομένων:

$$A = \begin{array}{c} \begin{matrix} M_1 & M_2 & M_3 \end{matrix} \\ \begin{bmatrix} 8 & 1 & 0 \\ 4 & 6 & 5 \\ 6 & 8 & 7 \\ 10 & 4 & 7 \\ 8 & 2 & 5 \\ 0 & 3 & 6 \end{bmatrix} \end{array} \begin{matrix} \text{Α} \\ \text{Β} \\ \text{Γ} \\ \Delta \\ \text{Ε} \\ \text{Ζ} \end{matrix}$$

Θα υπολογίσουμε τον πίνακα (συν) διακυμάνσεων και των πίνακα συσχετίσεων των μεταβλητών M_1, M_2, M_3.

(α) Πίνακας (συν)διακυμάνσεων

Οι μέσες τιμές των μεταβλητών (στηλών): $\bar{M}_1 = \dfrac{8+4+6+10+8+0}{6} = 6$,

$$\bar{M}_2 = \frac{1+6+8+4+2+3}{6} = 4, \qquad \bar{M}_3 = \frac{0+5+7+7+5+6}{6} = 5$$

Οι διασπορές των μεταβλητών M_1, M_2, M_3 είναι ίσες με:

$$\sigma^2_{M_1} = 10,67 \quad \sigma^2_{M_2} = 5,67 = \sigma^2_{M_3}$$

$M_1 - \bar{M}_1$	$M_2 - \bar{M}_2$	$(M_1 - \bar{M}_1)(M_2 - \bar{M}_2)$
8-6=2	-3	-6
-2	2	-4
0	4	0
4	0	0
2	-2	-4
-6	-1	6

$$-8 = \sum (M_1 - \bar{M}_1)(M_2 - \bar{M}_2)$$

Η συνδιασπορά των μεταβλητών M_1, M_2, είναι ίση με:

$$C(M_1,M_2) = \frac{1}{n}\sum (M_1 - \bar{M}_1)(M_2 - \bar{M}_2) = \frac{-8}{6} = -1,33$$

Ο πίνακας συνδιασποράς, (ανάλογοι υπολογισμοί) δίνεται από την σχέση:

$$C(M_1, M_2, M_3) = \begin{bmatrix} 10.67 & -1.33 & -1.33 \\ -1.33 & 5.67 & 3.67 \\ -1.33 & 3.67 & 5.67 \end{bmatrix}$$

Παρατηρείστε ότι ο πίνακας αυτός είναι συμμετρικός.

Για να πάρουμε τον πίνακα αυτό, από το δοσμένο πίνακα A, αφαιρούμε αρχικά από κάθε τιμή, τη μέση τιμή της αντίστοιχης στήλης. Έτσι προκύπτει ο πίνακας

$$A_0 = \begin{bmatrix} 8-6=2 & -3 & -5 \\ -2 & 2 & 0 \\ 0 & 4 & 2 \\ 4 & 0 & 2 \\ 2 & -2 & 0 \\ -6 & -1 & 1 \end{bmatrix}$$

("αποκλίσεις" σημείων από το *κέντρο βάρους*)

Ο ανάστροφος του πίνακα A_0 είναι ίσος με:

$$A_0^T = \begin{bmatrix} 2 & -2 & 0 & 4 & 2 & 6 \\ -3 & 2 & 4 & 0 & -2 & 1 \\ -5 & 0 & 2 & 2 & 0 & 1 \end{bmatrix}$$

Επομένως:

$$A_0^T A_0 = \begin{bmatrix} 64 & -8 & -8 \\ -8 & 34 & 22 \\ -8 & 22 & 34 \end{bmatrix}$$

Τότε, μπορούμε να παρατηρήσουμε ότι:

$$\left(\frac{1}{\sqrt{n}} A_0^T\right)\left(\frac{1}{\sqrt{n}} A_0\right) = \frac{1}{6} A_0^T A_0 = C(M_1, M_2, M_3)$$

Δηλαδή εάν συμβολίσουμε με $X = \frac{1}{\sqrt{n}} A_0$, οπότε $X^T = \frac{1}{\sqrt{n}} A_0^T$ τότε ο **πίνακας (συν)διακυμάνσεων** $C(M_1, M_2, M_3)$ των δεδομένων μας είναι ίσος με:

$$C(M_1, M_2, M_3) = X^T X$$

(β) Πίνακας συσχετίσεων

Ο πίνακας συσχετίσεων δίνεται από την σχέση:

$$\mathrm{C}_{\mathrm{ot}} = = \begin{bmatrix} \frac{64}{\sigma_{M_1}\sigma_{M_1}} & \frac{-8}{\sigma_{M_1}\sigma_{M_2}} & \frac{-8}{\sigma_{M_1}\sigma_{M_3}} \\ \frac{-8}{\sigma_{M_2}\sigma_{M_1}} & \frac{34}{\sigma_{M_2}\sigma_{M_2}} & \frac{22}{\sigma_{M_2}\sigma_{M_3}} \\ \frac{-8}{\sigma_{M_3}\sigma_{M_1}} & \frac{22}{\sigma_{M_3}\sigma_{M_2}} & \frac{34}{\sigma_{M_3}\sigma_{M_3}} \end{bmatrix} = \begin{bmatrix} 1 & -0{,}17 & -0{,}17 \\ -0{,}17 & 1 & 0{,}65 \\ -0{,}17 & 0{,}65 & 1 \end{bmatrix}$$

Αν θεωρήσουμε τον παρακάτω (κανονικοποιημένο πίνακα)

$$A_0^N = \begin{bmatrix} \frac{2}{\sigma_{M_1}} & \frac{-3}{\sigma_{M_2}} & \frac{-5}{\sigma_{M_3}} \\ \frac{-2}{\sigma_{M_1}} & \frac{2}{\sigma_{M_2}} & \frac{0}{\sigma_{M_3}} \\ \frac{0}{\sigma_{M_1}} & \frac{4}{\sigma_{M_2}} & \frac{2}{\sigma_{M_3}} \\ \frac{4}{\sigma_{M_1}} & \frac{0}{\sigma_{M_2}} & \frac{2}{\sigma_{M_3}} \\ \frac{2}{\sigma_{M_1}} & \frac{-2}{\sigma_{M_2}} & \frac{0}{\sigma_{M_3}} \\ \frac{6}{\sigma_{M_1}} & \frac{-1}{\sigma_{M_2}} & \frac{1}{\sigma_{M_3}} \end{bmatrix}$$

Ο ανάστροφός του είναι ίσος με:

$$\left(A_0^N\right)^T = \begin{bmatrix} \frac{2}{\sigma_{M_1}} & \frac{-2}{\sigma_{M_1}} & \frac{0}{\sigma_{M_1}} & \frac{4}{\sigma_{M_1}} & \frac{2}{\sigma_{M_1}} & \frac{-6}{\sigma_{M_1}} \\ \frac{-3}{\sigma_{M_2}} & \frac{2}{\sigma_{M_2}} & \frac{4}{\sigma_{M_2}} & \frac{0}{\sigma_{M_2}} & \frac{-2}{\sigma_{M_2}} & \frac{-1}{\sigma_{M_2}} \\ \frac{-5}{\sigma_{M_3}} & \frac{0}{\sigma_{M_3}} & \frac{2}{\sigma_{M_3}} & \frac{2}{\sigma_{M_3}} & \frac{0}{\sigma_{M_3}} & \frac{1}{\sigma_{M_3}} \end{bmatrix}$$

Τότε:

$$\left(\frac{1}{\sqrt{n}}\left(A_0^N\right)^T\right)\left(\frac{1}{\sqrt{n}}\right)A_0^N = \frac{1}{6}\left(A_0^N\right)^T A_0^N = C_{o\tau}$$

Δηλαδή εάν συμβολίσουμε με $X^N = \frac{1}{\sqrt{n}} A_0^N$, οπότε $\left(X^N\right)^T = \frac{1}{\sqrt{n}}\left(A_0^N\right)^T$ τότε ο **πίνακας συσχετίσεων** $C_{o\tau}$ των δεδομένων μας είναι ίσος με:

$$C_{o\tau} = \left(X^N\right)^T X^N$$

Παρατήρηση 2.4.2

Μερικοί συγγραφείς θεωρούν σαν πίνακα $X^N = \frac{1}{\sqrt{n-1}} A_0^N$. Αυτό οφείλεται στο πως ορίζει κανείς τη διασπορά μιάς τυχαίας μεταβλητής. Άλλοτε διαιρούμε με n και άλλοτε με n-1**.**

Β. Περιγραφή της μεθόδου

Για να υπολογίσει κάποιος τις κύριες συνιστώσες θα πρέπει να διαγωνοποιήσει τον **πίνακα συνδιασποράς/διακυμάνσεων** ή τον **πίνακα συσχετίσεων** των μεταβλητών (δηλαδή του πίνακα Χ).

Έστω ότι έχουμε p μεταβλητές/συνιστώσες, τις $\left(X_1\ X_2 ... X_p\right) = X$ και θέλουμε να δημιουργήσουμε νέες μεταβλητές $\left(Y_1\ Y_2 ... Y_p\right) = Y$ σα γραμμικό συνδυασμό των αρχικών μεταβλητών , με τη βοήθεια ενός πίνακα $U = \left(\upsilon_1\ \upsilon_2 ... \upsilon_p\right)$. Δηλαδή $Y = XU$. Για να υπολογίσουμε τις κύριες συνιστώσες θα πρέπει να υπολογίσουμε τον πίνακα U.

Παρατήρηση 2.4.3

Κάνουμε τις εξής παραδοχές:

(α) Κατά τον υπολογισμό των συνιστωσών θεωρούμε ότι αυτές είναι σε φθίνουσα σειρά ως προς το ποσοστό ερμηνείες της διακύμανσης των δεδομένων του αρχικού πίνακα.

(β) Θεωρούμε ότι:

$$\sum_{i=1}^{p} \upsilon_{ji}^2 = \upsilon_j^T \upsilon_j = 1, \quad j = 1, 2, ..., p$$

Για την <u>1η συνιστώσα</u> Y_1 έχουμε τα εξής: η διακύμανσή της θα είναι ίση με

$$Y_1^T Y_1 = \left(X\upsilon_1\right)^T X\upsilon_1 = \upsilon_1^T X^T X \upsilon_1 = \lambda_1 \upsilon_1^T \upsilon_1 = \lambda_1$$

Για να υπολογίσουμε δηλαδή το υ_1 θα πρέπει να μεγιστοποιήσουμε τη διακύμανση της Y_1 δηλαδή το $Y_1^T Y_1 = \upsilon_1^T X^T X \upsilon_1$ (με τον περιορισμό $\sum_{i=1}^{p} \upsilon_{ji}^2 = 1, \quad j = 1, 2, ..., p$), δηλαδή να μεγιστοποιήσουμε (με χρήση πολλαπλασιαστών Lagrange) τη συνάρτηση

$$L\left(\upsilon_1\right) = \upsilon_1^T X^T X \upsilon_1 - \lambda\left(\upsilon_1^T \upsilon_1 - 1\right)$$

απ' όπου έχουμε:

$$\frac{\partial L\left(\upsilon_1\right)}{\partial \upsilon_1} = 2\left(X^T X - \lambda I\right)\upsilon_1 = 0 \Rightarrow X^T X \upsilon_1 = \lambda \upsilon_1$$

Η τελευταία εξίσωση είναι η εξίσωση εύρεσης του ιδιοδιανύσματος υ_1 του πίνακα $X^T X$ με αντίστοιχη ιδιοτιμή λ_1 (παρατηρείστε ότι η διακύμανση της συνιστώσας Y_1 είναι ίση με λ_1).

Παρατήρηση 2.4.4

Για να κατασκευάσουμε τις κύριες συνιστώσες των δεδομένων μας (πίνακας X), θα πρέπει να βρούμε τις ιδιοτιμές και τα αντίστοιχα ιδιοδιανύσματα του πίνακα διακυμάνσεων $X^T X$ (ή του πίνακα των συσχετίσεων). Τότε:

(α) το ιδιοδιάνυσμα υ_1 που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή λ_1 είναι η 1^{η} κύρια συνιστώσα (διεύθυνση της). Το ιδιοδιάνυσμα υ_2 που αντιστοιχεί στη δεύτερη μεγαλύτερη ιδιοτιμή λ_2 είναι η 2^{η} κύρια συνιστώσα κ.ο.κ.

(β) η διακύμανση κάθε νέας συνιστώσας είναι ίση με την ιδιοτιμή που αντιστοιχεί στο αντίστοιχο ιδιοδιάνυσμα που την περιγράφει/ορίζει (δηλαδή $V\left(Y_i\right) = \lambda_i$).

(γ) οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους αφού ο πίνακας συσχετίσεων τους Δ όπως είδαμε παραπάνω είναι διαγώνιος.

(δ) Η ποσότητα:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

είναι το *ποσοστό της διακύμανσης των δεδομένων που ερμηνεύει η i-συνισττώσα*.

(ε) Η συνολική διακύμανση των κυρίων συνιστωσών ($tr\left(\Delta\right)$) είναι ίση με τη συνολική διακύμανση των αρχικών μεταβλητών ($tr\left(X^T X\right)$). Πράγματι:

$$tr\left(X^T X\right) = tr\left(U \Delta U^T\right) = tr\left(\Delta\right) = \lambda_1 + \lambda_2 + ... + \lambda_p$$

Αν χρησιμοποιήσουμε τον πίνακα συσχετίσεων τότε καθένα από τα διαγώνια στοιχεία του πίνακα $X^T X$ είναι ίσο με τη μονάδα, οπότε η συνολική διακύμανση είναι ίση με $tr\left(X^T X\right) = p$ (δηλαδή ίση με τον αριθμό των μεταβλητών).

(στ) (ιδιοτιμές και ιδιοδιανύσματα συμμετρικού πίνακα)

(i) Οι ιδιοτιμές ενός συμμετρικού πίνακα υπάρχουν πάντα. Είναι θετικοί αριθμοί ή μηδέν

(ii) Τα ιδοδιανύσματα που αντιστοιχούν σε διαφορετικές ιδιοτιμές είναι ορθογώνια. Οι συντεταγμένες τους είναι πραγματικοί αριθμοί.

(iii) Τα ιδιοδιανύσματα αυτά μπορούν να θεωρηθούν στήλες ενός πίνακα U ο οποίος είναι ορθογώνιος ($U^{-1} = U^T$)

Παρατήρηση 2.4.5 (ειδικές περιπτώσεις Α.Κ.Σ.)

1) Αν μια μεταβλητή, είναι ασυσχέτιστη με τις υπόλοιπες, τότε κάποια κύρια συνιστώσα θα ταυτιστεί μαζί της. Σαν ασυσχέτιστη με τις άλλες δεν συνεισφέρει στατιστικά στη μελέτη του προβλήματος του ενδιαφέροντός μας, και συνήθως αφαιρείται (δεν έχει νόημα να τη συμπεριλάβουμε).

2) Αν έχουμε μηδενικές ιδιοτιμές, σημαίνει ότι ο πίνακας της ανάλυσης (πίνακας διακυμάνσεων ή πίνακας συσχετίσεων) δεν είναι πλήρους βαθμού ή με άλλα λόγια μερικές μεταβλητές είναι εξαρτημένες. Συνήθως τις παραλείπουμε. Στην πράξη δεν βρίσκουμε μηδενικές ιδιοτιμίες αλλά σχεδόν μηδενικές. Σε τέτοιες ιδιοτιμές αντιστοιχούν συνιστώσες με μηδενική ερμηνεία διακύμανσης (εξαρτημένες μεταβλητές παρέχουν ίδια πληροφορία).

3) Αν δυο ιδιοτιμές είναι ίδιες, τότε σε αυτές αντιστοιχούν ίδιες κύριες συνιστώσες (πλεονασμός). Στην πράξη, με δεδομένα από δείγμα, είναι σπάνιο να συμβεί. Αν συμβαίνει, υπάρχει κάποιο πρόβλημα στα δεδομένα (επανάληψη στηλών).

4) Μπορεί να συμβεί, δύο ίδια ιδιοδιανύσματα να αντιστοιχούν σε διαφορετικές ιδιοτιμές. Τότε παίρνουμε τις ίδιες συνιστώσες αλλά αντιστοιχούν σε διαφορετικό ποσοστό ερμηνείας της διακύμανσης.

5) Αν ο πίνακας συσχετίσεων έχει μόνο θετικά στοιχεία (συσχετίσεις), η 1^{η} κύρια συνιστώσα είναι ο σταθμικός μέσος όρος των αρχικών μεταβλητών με σταθμίσεις τους αντίστοιχους συντελεστές (συντεταγμένες ιδιοδιανύσματος που αντιστοιχεί στη μέγιστη ιδιοτιμή).

Παράδειγμα 2.4.6

Έστω ότι ο πίνακας διακυμάνσεων, ενός συνόλου δεδομένων, είναι ο

$$X = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Οι ιδιοτιμές του πίνακα είναι οι (κατά φθίνουσα σειρά):

$$\lambda_1 = 3+\sqrt{2},\ \lambda_2 = 3,\ \lambda_3 = 3-\sqrt{2}$$

τα δε αντίστοιχα ιδιοδιανύσματα είναι:

$$\upsilon_1 = \left(3+\sqrt{2},1,0\right),\ \upsilon_2 = (0,0,1),\ \upsilon_3 = \left(3-\sqrt{2},1,0\right)$$

Με τη βοήθεια των ιδιοδιανυσμάτων, οι κύριες συνιστώσες είναι οι:

$$Y_1 = \left(3+\sqrt{2}\right)X_1 + X_2 + 0X_3 = \left(3+\sqrt{2}\right)X_1 + X_2$$
$$Y_2 = \left(3+\sqrt{2}\right)X_1 + X_2 + 0X_3 = X_3$$
$$Y_3 = \left(3-\sqrt{2}\right)X_1 + X_2 + 0X_3 = \left(3-\sqrt{2}\right)X_1 + X_2$$

Παρατήρηση 2.4.7

Στο παραπάνω παράδειγμα, όπως μπορεί να δει κανείς από τον πίνακα διακυμάνσεων, η μεταβλητή X_3 είναι ασυσχέτιστη με τις X_1, X_2. Επομένως περιμένει κανείς η X_3 να είναι μια από τις κύριες συνιστώσες (εδώ η Y_2). Γι' αυτό συνήθως απαλείφουμε τη μεταβλητή από τα δεδομένα μας.

Παράδειγμα 2.4.6 (συνέχεια)

Οι διασπορές των κύριων συνιστωσών είναι:

$$V\left(Y_1\right) = \lambda_1 = 3+\sqrt{2}$$
$$V\left(Y_2\right) = \lambda_2 = 3,$$
$$V\left(Y_3\right) = \lambda_3 = 3-\sqrt{2}$$

Η 1^{η} συνιστώσα ερμηνεύει το: $\dfrac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = \dfrac{3+\sqrt{2}}{3+\sqrt{2}+3+3-\sqrt{2}} = \dfrac{3+\sqrt{2}}{9} = 0,49$

δηλαδή 49% της διακύμανσης των δεδομένων. Η 2^{η} συνιστώσα ερμηνεύει το: $\dfrac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3} = \dfrac{3}{9} = 0,33$ δηλαδή 33% της διακύμανσης των δεδομένων. Η 3^{η}

συνιστώσα ερμηνεύει το: $\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}=\frac{3-\sqrt{2}}{9}=0,18$ δηλαδή 18% της διακύμανσης των δεδομένων. Παρατηρείστε ότι οι δύο πρώτες κύριες συνιστώσες ερμηνεύουν το 82% της διακύμανσης (οπότε είναι αρκετές για να παραστήσουν ικανοποιητικά τα δεδομένα.

Παράδειγμα 2.4.8

(α) Έστω ότι ο πίνακας διακυμάνσεων, ενός συνόλου δεδομένων, είναι ο

$$X=\begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

Οι ιδιοτιμές του πίνακα είναι οι (κατά φθίνουσα σειρά):

$$\lambda_1=\frac{3+\sqrt{5}}{2},\quad \lambda_2=\frac{3-\sqrt{5}}{2}$$

Η 1^{η} συνιστώσα ερμηνεύει το: $\frac{\lambda_1}{\lambda_1+\lambda_2}=\frac{\frac{3+\sqrt{5}}{2}}{\frac{3+\sqrt{5}}{2}+\frac{3-\sqrt{5}}{2}}=\frac{3+\sqrt{5}}{6}=0,87$ δηλαδή το 87% της διακύμανσης των δεδομένων. Η 2^{η} συνιστώσα ερμηνεύει το υπόλοιπο 13% της διακύμανσης των δεδομένων.

(β) Έστω ότι ο πίνακας διακυμάνσεων, ενός συνόλου δεδομένων, είναι ο

$$X=\begin{bmatrix} 200 & -1 \\ -1 & 1 \end{bmatrix}$$

Οι ιδιοτιμές του πίνακα είναι οι (κατά φθίνουσα σειρά):

$$\lambda_1=\frac{201+89\sqrt{5}}{2},\quad \lambda_2=\frac{201-89\sqrt{5}}{2}$$

Η 1^{η} συνιστώσα ερμηνεύει το:

$$\frac{\lambda_1}{\lambda_1+\lambda_2}=\frac{\frac{201+89\sqrt{5}}{2}}{\frac{201+89\sqrt{5}}{2}+\frac{201-89\sqrt{5}}{2}}=\frac{201+89\sqrt{5}}{402};\ 0,995$$

δηλαδή το 99,5% της διακύμανσης των δεδομένων. Η 2^{η} συνιστώσα ερμηνεύει το υπόλοιπο 0,5% της διακύμανσης των δεδομένων.

Παρατήρηση 2.4.9

Αν παρατηρήσει κανείς τα παραδείγματα 2.5.8(α) και 2.5.8(β) , η μόνη διαφορά στον πίνακα διακυμάνσεων είναι ότι η διασπορά της X_1, στο (β) παράδειγμα, είναι πολύ μεγάλη. Αποτέλεσμα, η 1^{η} συνιστώσα ερμηνεύει σχεδόν όλη τη διακύμανση των δεδομένων. Στη περίπτωση αυτή προτιμότερο θα ήταν να χρησιμοποιηθεί ο πίνακας των συσχετίσεων.

(γ) Έστω ότι ο πίνακας διακυμάνσεων, ενός συνόλου δεδομένων, είναι ο

$$X = \begin{bmatrix} 2 & -0,01 \\ -0,01 & 0,001 \end{bmatrix}$$

Οι ιδιοτιμές του πίνακα είναι οι (κατά φθίνουσα σειρά):

$$\lambda_1 = 2,005, \quad \lambda_2 = 0,0005$$

Η 1^{η} συνιστώσα ερμηνεύει το: $\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{2,005}{2,005 + 0,0005} = 0,9997$ δηλαδή 99,97% της διακύμανσης των δεδομένων. Η 2^{η} συνιστώσα ερμηνεύει το υπόλοιπο 0,3% της διακύμανσης των δεδομένων.

Παρατήρηση 2.4.10

Εάν υποθέσουμε ότι στο παράδειγμα 2.5.8 (α) η μεταβλητή X_1 παριστάνει το βάρος (σε kg) και η μεταβλητή X_2 το μήκος αντικειμένου (σε dm). Τότε, στο παράδειγμα 2.5.8 (γ) η μεταβλητή X_1 μπορεί να παριστάνει το βάρος (σε kg) και η μεταβλητή X_2 το μήκος αντικειμένου (σε m). Αλλάζουμε δηλαδή τη μονάδα μέτρησης της μεταβλητής. Το αποτέλεσμα είναι ότι αλλάζει το ποσοστό ερμηνείας της διακύμανσης από κάθε μια από τις κύριες συνιστώσες (εδώ η 1^{η} συνιστώσα ερμηνεύει σχεδόν όλη τη διακύμανση των δεομένων). Στη περίπτωση αυτή προτιμότερο θα ήταν να χρησιμοποιηθεί ο πίνακας των συσχετίσεων.

2.5 Πίνακας (συν)διακυμάνσεων ή πίνακας συσχετίσεων

Ένα ακόμα σημαντικό πρόβλημα, όπως είδαμε και στα παραδείγματα, είναι το ποιόν πίνακα θα χρησιμοποιήσει κανείς για να υπολογίσει τις κύριες συνιστώσες των δεδομένων μας. Τον πίνακα (συν)διακυμάνσεων ή τον πίνακα συσχετίσεων. Αν χρησιμοποιήσει τον πίνακα διακυμάνσεων οι κύριες συνιστώσες «επηρεάζονται» από

τις μονάδες μέτρησης των μεταβλητών και από το μέγεθος της διασποράς κάθε μιάς από τις αρχικές μεταβλητές. Γι' αυτό χρησιμοποιούμε συνήθως τον πίνακα συσχετίσεων. Οι συσχετίσεις δεν αλλάζουν όταν αλλάζει η κλίμακα ή οι μονάδες μέτρησης των μεταβλητών.

Μερικές φορές όμως η χρήση του πίνακα συσχετίσεων δεν ενδείκνυται, επειδή η διασπορά στις διακυμάνσεις ίσως να περιέχει πληροφορία πολύτιμη για το πρόβλημα που μελετάμε.

Άρα στην πράξη δεν είναι φανερό ποιό από τους δυο αυτούς πίνακες θα πρέπει να χρησιμοποιούμε. Θα πρέπει ίσως να αποφεύγουμε να χρησιμοποιούμε τον πίνακα διακύμανσης εάν υπάρχουν μεταβλητές με πολύ μεγαλύτερη διακύμανση από τις υπόλοιπες (αν διαφέρουν ως προς τη διακύμανση αλλά αναφέρονται στις ίδιες μονάδες, τότε θα είναι καλό να χρησιμοποιούμε τον πίνακα).

2.6 Χρήση Ανάλυσης Κύριων Συνιστωσών

1) **Πολλαπλή (γραμμική) παλινδρόμηση**: συνήθως υπάρχει το πρόβλημα της πολυσυγγραμμικότητας (συμβαίνει όταν οι ανεξάρτητες μεταβλητές είναι συσχετισμένες). Αποτέλεσμα: όχι συνεπείς εκτιμήτριες συναρτήσεις ελαχίστων τετραγώνων, μεγάλες διακυμάνσεις και πρόσημα των συντελεστών χωρίς ερμηνεία. Χρησιμοποιώντας μέρος των κυρίων συνιστωσών (ασυσχέτιστες μεταβλητές) το πρόβλημα αποφεύγεται. Εδώ δεν μπορεί να ερμηνεύσει κανείς τους συντελεστές (μερικές φορές υπάρχει ερμηνεία σε εικονομετρικές εφαρμογές/ποσοτικοποίηση αφηρημένων εννοιών).

2) Τα πολυδιάστατα δεδομένα είναι δύσκολο να παρασταθούν γραφικά. Αν παραστήσουμε τα δεδομένα αναφορικά με τις (πρώτες 2ή 3) κύριες συνιστώσες έχουμε οπτική παρουσίαση των δεδομένων.

3) Από τα σκορ (συντεταγμένες) των παρατηρήσεων στις κύριες συνιστώσες μπορεί να παρατηρήσει κάποιος πως ομαδοποιούνται οι παρατηρήσεις.

4) **Έλεγχος ποιότητας**: πολλές φορές παρατηρούμε αρκετά χαρακτηριστικά ενός προϊόντος. Με τη χρήση διαγραμμάτων ελέγχου είναι δύσκολο να δει κανείς αν η

διαδικασία παραγωγής του προϊόντος είναι εκτός ελέγχου. Μειώνοντας με την Α.Κ.Σ. τις μεταβλητές (διαστάσεις), το πρόβλημα ίσως γίνεται ευκολότερο.

5) **Data mining**: θέλουμε να κάνουμε στατιστική ανάλυση σε πολύ μεγάλες βάσεις δεδομένων. Περιορίζοντας τις διαστάσεις με Α.Κ.Σ., το πρόβλημα γίνεται ευκολότερο.

2.7 Αριθμός κύριων συνιστωσών

Ένα από τα προβλήματα που αντιμετωπίζει κανείς στην ανάλυση κύριων συνιστωσών είναι ο προσδιορισμός του αριθμού των κύριων συνιστωσών που θα επιλέξει για να περιγράψει τα δεδομένα (δεν επιλέγονται όλες οι κύριες συνιστώσες αφού όπως αναφέραμε σκοπός της εν λόγω ανάλυσης είναι ο περιορισμός των διαστάσεων του προβλήματος). Η απάντηση εδώ δεν είναι μονοσήμαντη. Αναφέρουμε παρακάτω τα βασικά κριτήρια που χρησιμοποιούνται για την επιλογή του αριθμού των κύριων συνιστωσών.

1) Ποσοστό της συνολικής διακύμανσης που εξηγούν οι συνιστώσες

Βάζουμε ένα όριο στο ποσοστό της διασποράς των δεδομένων που θα πρέπει να ερμηνεύουν οι νέες συνιστώσες. Το όριο αυτό είναι συνήθως υψηλό (π.χ. 80%). Διαλέγουμε λοιπόν τόσες συνιστώσες ώστε αθροιστικά να εξηγούν ποσοστό διασποράς μεγαλύτερο (ή περίπου ίσο) από αυτό που βάλαμε σαν όριο.

Μερικές φορές, ιδιαίτερα αν ο στόχος είναι υψηλός, δεν παίρνουμε τα καλύτερα δυνατά αποτελέσματα αφού είναι απαραίτητο (για να επιτευχθεί ο στόχος) να χρησιμοποιήσουμε μεγάλο αριθμό κύριων συνιστωσών

2) Κριτήριο Kaiser

Εάν $\lambda_1, \lambda_2, ..., \lambda_p$ είναι οι ιδιοτιμές του πίνακα συνδιαποσράς ή συσχετίσεων και $\overline{\lambda} = \frac{\sum_{j=1}^{p} \lambda_j}{p}$ η μέση τους τιμή, τότε επιλέγουμε τόσες συνιστώσες όσος και ο αριθμός των $\lambda_j > \overline{\lambda}$.

Εάν χρησιμοποιήσουμε τον πίνακα συσχετίσεων τότε $\overline{\lambda}=1$. Άρα επιλέγουμε τόσες συνιστώσες όσος και ο αριθμός των $\lambda_j>1$.

Με το κριτήριο αυτό γίνεται συνήθως υπερεκτίμηση του αριθμού των κυρίων συνιστωσών

3) Ποσοστό διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται

Επιλέγουμε τόσες συνιστώσες ώστε για κάθε μεταβλητή να ερμηνεύεται ένα μεγάλο ποσοστό της διακύμανσης/διασποράς της (το ποσοστό αυτό είναι υποκειμενικό, εξ ού και το μειονέκτημα της μεθόδου)

4) Scree plot

Κατασκευάζουμε ένα γράφημα (το ονομαζόμενο **scree plot**) όπου στο x-άξονα τοποθετούμε ένα αύξοντα αριθμό/σειρά και στον y-άξονα τις τιμές των ιδιοτιμών λ_j. Στο γράφημα τοποθετούμε επίσης τα σημεία $\left(j,\lambda_j\right),\ j=1,2,...,p$ και τα ενώνουμε με μια τεθλασμένη γραμμή. Διαλέγουμε τόσες συνιστώσες όσος και ο αριθμός των j, ώσπου η γραμμή που συνδέει τα $\left(j,\lambda_j\right),\ j=1,2,...,p$ να γίνει οριζόντια. Το μειονέκτημα της μεθόδου είναι ότι, πολλές φορές είναι υποκειμενικός ο τρόπος που εκτιμά κανείς το σημείο που η γραμμή γίνεται οριζόντια.

5) Μέθοδος σπασμένου ραβδιού (broken stick)

Αν σπάσουμε ένα ραβδί σε p κομμάτια, το κ μεγαλύτερο θα έχει μήκος:

$$g_k=\frac{\sum_{j=1}^{p}\left(\frac{1}{j}\right)}{p}$$

Επιλέγουμε τόσα k ώστε: $\frac{\lambda_k}{\sum_{j=1}^{p}\lambda_j}>g_k$.

2.8 Βήματα της μεθόδου

Αναφέρουμε παρακάτω τα βήματα που ακολουθεί κανείς για τον υπολογισμό των κυρίων συνιστωσών ενός συνόλου δεδομένων.

ΒΗΜΑ 1 Λήψη δεδομένων

Χρησιμοποιούμε όπως αναφέραμε έναν αρχικό πίνακα δεδομένων A (n,p), με n παρατηρήσεις και p μεταβλητές. Η τιμή που εμφανίζεται στον πίνακα δεδομένων για τη μεταβλητή j στην παρατήρηση i συμβολίζεται με a_{ij}.

Θεωρούμε την κάθε γραμμή i του πίνακα A, δηλαδή κάθε μια από τις n παρατηρήσεις σαν ένα σημείο (διάνυσμα) στο χώρο των p διαστάσεων. Οι τιμές που παίρνει η παρατήρηση i ως προς κάθε μια από τις p μεταβλητές, θεωρούνται ως οι συντεταγμένες του αντίστοιχου σημείου στους p άξονες αυτού του χώρου. Το σύνολο λοιπόν των πληροφοριών που μας παρέχει ο πίνακας των δεδομένων A, μπορεί να παρομοιαστεί με **ένα νέφος** n σημείων στο χώρο $\mathbf{R}^p$.

Σκοπός μας να καταφέρουμε να παρατηρήσουμε το νέφος των σημείων σε ένα χώρο πολύ λιγότερων διαστάσεων από τις αρχικές, δηλαδή m (m<P). Για να το επιτύχουμε αυτό προσπαθούμε να ορίσουμε ένα γραμμικό συνδυασμό των αρχικών μεταβλητών (διευθύνσεις αξόνων), m διαστάσεων που να διέρχεται όσο το δυνατόν πλησιέστερα από το κέντρο μάζας του αρχικού νέφους των σημείων των δεδομένων με την παρακάτω έννοια.

ΒΗΜΑ 2 Κανονικοποίηση μεταβλητών

Πολλές φορές τα στοιχεία του πίνακα A είναι στοιχεία τελείως ανομοιογενή και δεν είναι αριθμοί που εκφράζουν συχνότητα. Έχουμε δηλαδή p μεταβλητές με διαφορετικές μονάδες η κάθε μια. Για να μπορέσουμε να μελετήσουμε αυτόν τον πίνακα δεδομένων, θα πρέπει να κανονικοποιήσουμε τις τιμές της κάθε μεταβλητής αφαιρώντας από αυτές την αντίστοιχη μέση της τιμή και διαιρώντας τη διαφορά με την τυπική της απόκλιση σ_X (της μεταβλητής) .

Παρατήρηση 2.8.1

Δύο βασικές ιδιότητες που έχει ο πίνακας των κανονικοποιημένων δεδομένων είναι:

- Για οποιαδήποτε από τις μεταβλητές (στήλες) το άθροισμα των στοιχείων είναι 0.
- Το άθροισμα των τετραγώνων όλων των στοιχείων του πίνακα ισούται με το πλήθος των παρατηρήσεων (γραμμών) n.

ΒΗΜΑ 3 Υπολογισμός του πίνακα (συν)διακύμανσης και του πίνακα συσχέτισης των αρχικών μεταβλητών.

Δημιουργούμε τον τετραγωνικό πίνακα V (συν)**διακύμανσης** ο οποίος ισούται με το γινόμενο του αρχικού «κεντρικοποιημένου» πίνακα (δηλαδή του πίνακα που προκύπτει από τον αρχικό πίνακα δεδομένων αν αφαιρέσουμε από τις τιμές κάθε στήλης τη μέση τιμή της στήλης) επί τον ανάστροφο του. Δημιουργούμε επίσης τον τετραγωνικό πίνακα **συσχέτισης R** ο οποίος είναι ίσος με το γινόμενο του αρχικού «κανονικοποιημένου» πίνακα επί τον ανάστροφο του πολλαπλασιασμένο με $\frac{1}{n-1}$.

Παρατήρηση 2.8.2

Όπως είδαμε σε προηγούμενες σελίδες το γινόμενο ενός πίνακα επί τον ανάστροφο του μας δίνει έναν τετραγωνικό πίνακα.

Στην ανάλυση κυρίων συνιστωσών το γινόμενο $X^T X$ μας δίνει έναν πίνακα διαστάσεων (p,p) ($\mathrm{X^T X} \to (\rho, n)\times(n, \rho) = (\rho, \rho)$), ενώ το XX^T μας δίνει ένα πίνακα διαστάσεων: $(n, \rho)\times(\rho, n) = (n, n)$.

Συνήθως χρησιμοποιούμε το γινόμενο $X^T X$ (ο αριθμός των μεταβλητών είναι μικρότερος από τον αριθμό των περιπτώσεων). Άρα

$$R = \frac{1}{n-1} X^T X$$

ΒΗΜΑ 4 Υπολογισμός των ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα συσχέτισης

Υπολογίζουμε τις ιδιοτιμές και τα ιδιοδιανύσματα του **πίνακα συσχέτισης** R**.** Για τον υπολογισμό των ιδιοτιμών χρησιμοποιούμε την εξίσωση $\det\left(R - \lambda I\right) = 0$ όπου I ο μοναδιαίος πίνακας pxp.

Για κάθε ιδιοτιμή, υπολογίζουμε το αντίστοιχο ιδιοδιάνυσμα (τα ιδιοδιανύσματα είναι οι διευθύνσεις των κυρίων συνιστωσών -- **άξονες αδράνειας).**

Αφού υπολογιστούν τα ιδιοδιανύσματα του πίνακα συσχέτισης, τα τοποθετούμε σε μία σειρά σύμφωνα με τις αντίστοιχες τιμές των ιδιοτιμών θεωρούμενες από τη μεγαλύτερη στη μικρότερη ($\lambda_1 > \lambda_2 > ... > \lambda_p$).

ΒΗΜΑ 5 Επιλογή αριθμού κυρίων συνιστωσών

Από τις p ιδιοτιμές – ιδιοδιανύσματα μόνο ορισμένες όπως έχουμε αναφέρει προσφέρουν σημαντική πληροφορία (m<p). Πρόκειται για τις **κύριες συνιστώσες.**

Μερικά κριτήρια για την επιλογή του αριθμού των κυρίων συνιστωσών αναφέρθησαν παραπάνω. Αναφέρουμε και εδώ μερικά περιληπτικά

1ο. Σύμφωνα με το **κριτήριο Guttmann & Kaiser** κρατάμε όσα ιδιοδιανύσματα έχουν αντίστοιχη ιδιοτιμή μεγαλύτερη από 1.

2ο. Οι συνιστώσες που αντιστοιχούν σε μικρή ιδιοτιμή λ_j και ταυτόχρονα δεν διαφέρουν μεταξύ τους σημαντικά (δηλαδή) λ_j ; λ_{j+1} ; λ_{j+2} κ.α. δεν αποτελούν κυρίες συνιστώσες (Cattell)

(χρησιμοποιούμε το Scree plot: διάγραμμα που περιγράφει τη σχέση μεταξύ του αριθμού συνιστωσών και την τιμή των ιδιοτιμών).

3ο. Κρατάμε τόσες συνιστώσες όσες να ερμηνεύουν αθροιστικά το 70-80% της συνολικής διακύμανσης. (το ποσοστό της διακύμανσης των δεδομένων που ερμηνεύει η j συνιστώσα είναι ίσο με $\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$).

Στη διπλωματική εργασία θα χρησιμοποιήσουμε το κριτήριο αυτό.

Δημιουργούμε μια σειρά από νέες μεταβλητές Z_j, $j = 1, 2, ..., p$ (σαν γραμμικούς συνδυασμούς των αρχικών μεταβλητών). Οι διακυμάνσεις (μεταβλητότητα) που αναπτύσσονται μεταξύ των νέων μεταβλητών, διαβαθμίζονται με τέτοιο τρόπο ώστε η πρώτη μεταβλητή Z_1 επιλέγεται να εξηγεί ένα όσο το δυνατόν μέγιστο ποσοστό της ολικής μεταβλητότητας των αρχικών δεδομένων, Z_2, ένα δεύτερο μέγιστο ποσοστό της διακύμανσης .

Οι νέες μεταβλητές Z_j, $j=1,2,...,m$ είναι οι κύριες συνιστώσες και με τον τρόπο αυτό δημιουργούνται ολιγάριθμες (m το πλήθος) συνιστώσες, οι οποίες ωστόσο εξηγούν μεγάλο ποσοστό της ολικής αδράνειας $\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i}$.

Παρατήρηση 2.8.3

Οι υπόλοιπες δευτερεύουσες συνιστώσες εξηγούν μικρό έως και ελάχιστο ποσοστό της διακύμανσης των δεδομένων και συνεπώς το στατιστικό τους αποτέλεσμα μπορεί να αγνοηθεί χωρίς την απώλεια ουσιαστικής πληροφορίας.

Για να είμαστε ακριβείς εάν αρχικά είχαμε p διαστάσεις στα δεδομένα μας και υπολογίσουμε p ιδιοτιμές, ιδιοδιανύσματα και στη συνέχεια επιλέξουμε μόνο m ιδιοδιανύσματα από τα αρχικά τότε τα τελικά μας δεδομένα θα έχουν m διαστάσεις.

ΒΗΜΑ 6 Υπολογισμός συντεταγμένων δεδομένων στις κύριες συνιστώσες

Θεωρούμε σαν πίνακα U τον πίνακα που έχει σαν j στήλη τις συντεταγμένες του ιδιαοδιανύσματος που αντιστοιχεί στην ιδιοτιμή λ_j. Εάν πολλαπλασιάσουμε (από δεξιά) τον A με τον πίνακα U, προκύπτει ο πίνακας των συντεταγμένων των δεδομένων μας ως προς τους νέους άξονες (κύριες συνιστώσες).

2.9 Αναλυτικό παράδειγμα Ανάλυσης Κυρίων Συνιστωσών

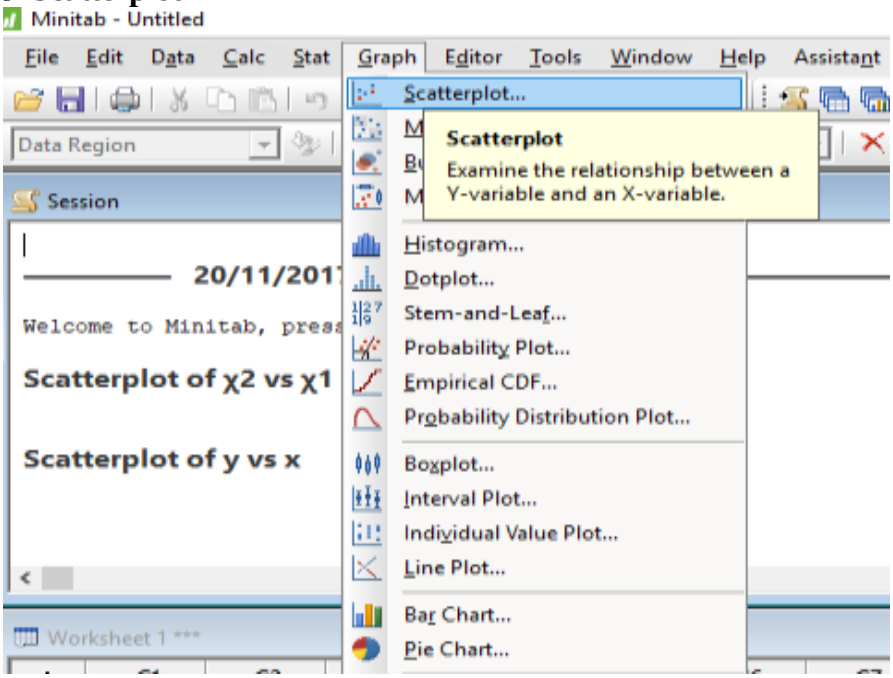
Έστω ότι έχουμε τις παρακάτω 2 μεταβλητές και 10 παρατηρήσεις.

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2.0	1.6
1.0	1.1
1.5	1.6
1.1	0.9

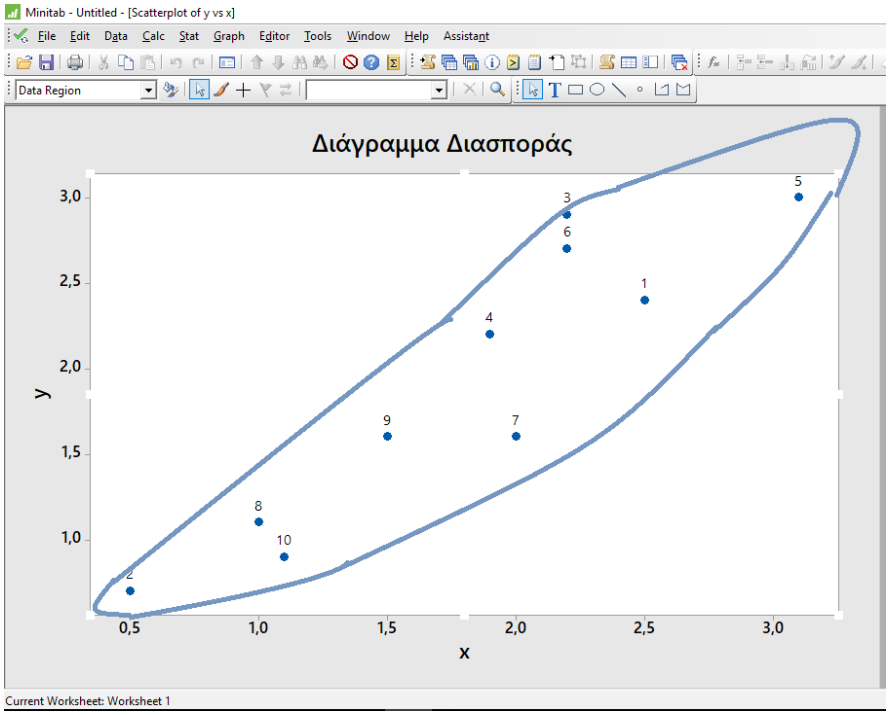
Θα κατασκευάσουμε αρχικά το διάγραμμα διασποράς για να δούμε πως κατανέμονται τα δεδομένα. Με τον τρόπο αυτό θα δούμε αν τα δεδομένα έχουν θετική ή αρνητική – ισχυρή ή ασθενή συσχέτιση.

Σε δύο στήλες του *Minitab* καταγράφουμε τα δεδομένα. Από το μενού επιλέγουμε:

Graph → Scatterplot



Παίρνουμε το παρακάτω διάγραμμα διασποράς



Στην συγκεκριμένη περίπτωση παρατηρούμε ***ισχυρή θετική συσχέτιση*** (θετική: τα δεδομένα έχουν μια αυξητική τάση // ισχυρή γιατί η αύξηση μεταξύ των μεταβλητών Χ και Υ είναι ανάλογη).

Παρατηρούμε ακόμα ότι το κέντρο βάρους των σημείων δεν είναι στο (0,0), επομένως θα μετατοπίσουμε/μετασχηματίσουμε τα δεδομένα.

ΒΗΜΑ 1 Λήψη δεδομένων // Κανονικοποίηση μεταβλητών

Θα κανονικοποιήσουμε τα δεδομένα: θα αφαιρέσουμε από κάθε μια από τις παρατηρήσεις τη μέση τιμή της αντίστοιχης μεταβλητής και θα διαιρέσουμε με τη τυπική απόκλισή της. Θα δημιουργήσουμε τις νέες μεταβλητές x*,y*:

excel: για τις μέσες τιμές χρησιμοποιούμε τις συναρτήσεις

$\bar{X}$ **= AVERAGE(στήλη x)**

$\bar{Y}$ **=AVERAGE(στήλη y)**

για τις τυπικές αποκλίσεις χρησιμοποιού τις συναρτήσεις

σ_X **= STDEV(στήλη χ)**

σ_y **=STDEV(στήλη y)**

F	G
x*	y*
0,879593097	0,578856809
-1,633530037	-1,429421916
0,502624627	1,169527022
0,125656157	0,342588724
1,633530037	1,287661065
0,628280783	0,933258937
0,251312313	-0,366215532
-1,005249254	-0,956885746
-0,37696847	-0,366215532
-1,005249254	-1,193153831

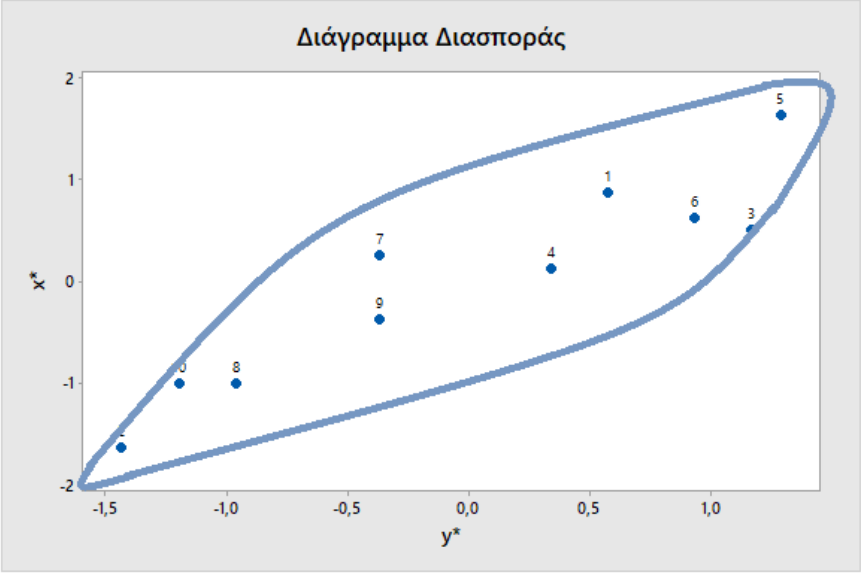
Παρατηρούμε ότι ικανοποιείται και η βασική ιδιότητα του κανονικοποιημένου πίνακα, δηλ το άθροισμα των στοιχείων κάθε στήλης είναι 0.

- Χρησιμοποιώντας στο excel τη συνάρτηση **sum(στήλη)** μας δίνει ότι το άθροισμα κάθε στήλης να είναι 0.

=SUM(F2:F11)

F	G
x*	y*
0,879593097	0,578856809
-1,633530037	-1,429421916
0,502624627	1,169527022
0,125656157	0,342588724
1,633530037	1,287661065
0,628280783	0,933258937
0,251312313	-0,366215532
-1,005249254	-0,956885746
-0,37696847	-0,366215532
-1,005249254	-1,193153831

Παρακάτω δίνουμε το διάγραμμα διασποράς των κανονικοποιημένων δεδομένων (ακολουθώντας τις ίδιες εντολές με πριν στο minitab) είναι:



Βλέπουμε ότι το κέντρο πλέον είναι στο (0,0) . Κάθε παρατήρηση παρουσιάζει μια απόκλιση από το μέσο σημείο, έτσι το σύνολο των παρατηρήσεων θα ορίζει ένα **νέφος σημείων** με κέντρο το μέσο σημείο 0.

ΒΗΜΑ 2 Υπολογισμός του πίνακα συσχετίσεων

Έστω ο πίνακας

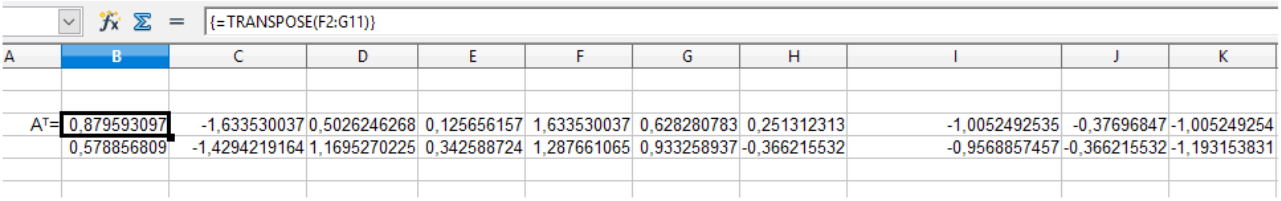
$$A=\begin{bmatrix} 0.879593 & 0.578857 \\ -1.63353 & -1.42942 \\ 0.502625 & 1.16953 \\ 0.125656 & 0.342589 \\ 1.63353 & 1.28766 \\ 0.628281 & 0.933259 \\ 0.251312 & -0.366216 \\ -1.00525 & -0.956886 \\ -0.376968 & -0.366216 \\ -1.00525 & -1.19315 \end{bmatrix}$$

Υπολογίζω τον πίνακα συσχέτισης.

Ο πίνακας Α βλέπουμε ότι έχει διάσταση 10x2 επομένως ο πίνακας A^T θα έχει διάσταση 2x10. **Άρα το γινόμενο A^TA θα έχει διάσταση**: (2x10) x (10x2) =(2,2)

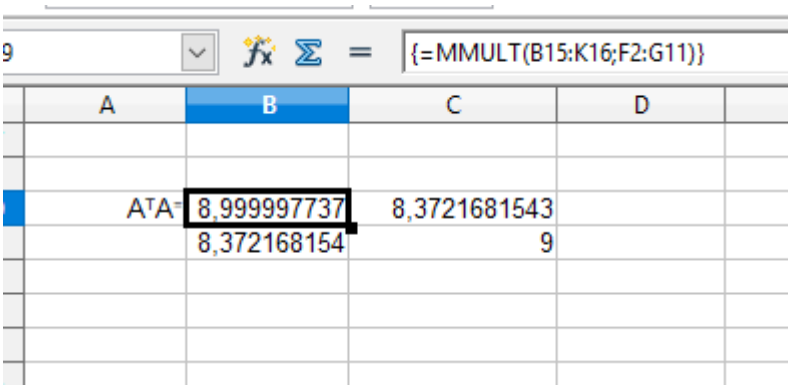
excel: Ο ανάστροφος υπολογίζεται με την εντολή

TRANSPOSE(πίνακας Α)



Το γινόμενο των πινάκων A^TA υπολογίζεται με την εντολή

MMULT(πίνακας A^T; πίνακας Α)

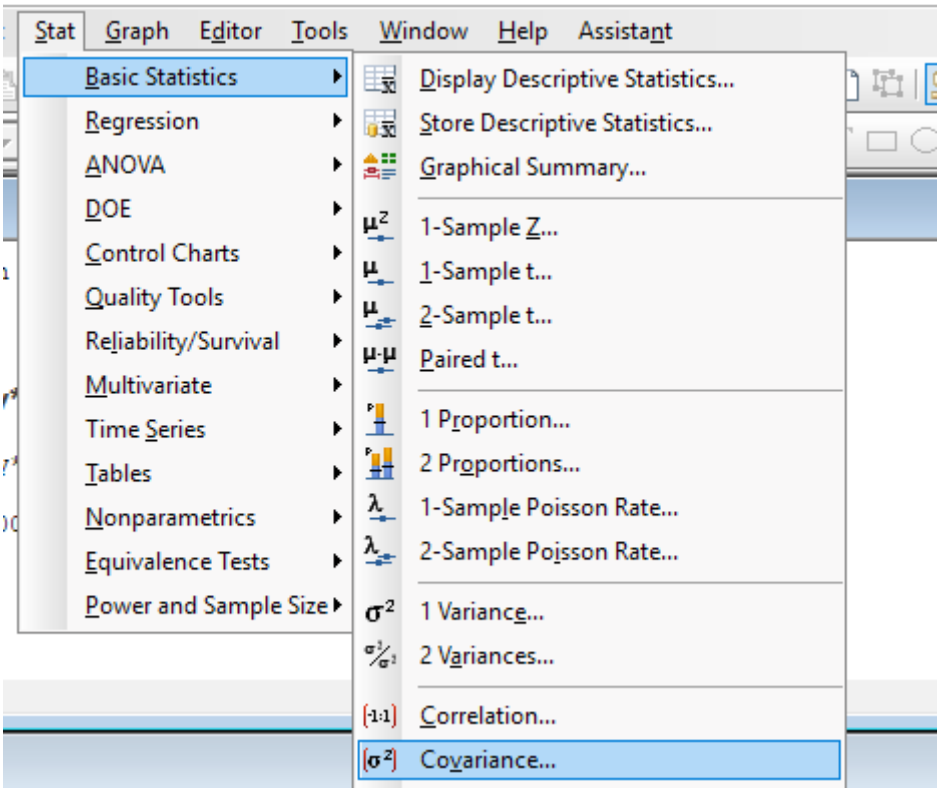


Ο πίνακας συσχετίσεων των δεδομένων μας είναι ίσος με:

C=	0,999999749	0,930240906
	0,930240906	1

Θα επαληθεύσουμε τα παραπάνω με τη βοήθεια του minitab.

Minitab εντολές **stat → basic stats → covariance**



Εισάγουμε τις κανονικοποιημένες μεταβλητές. Μας εμφανίζεται ο πίνακας σσυσχετίσεων:

Covariances: x*; y*

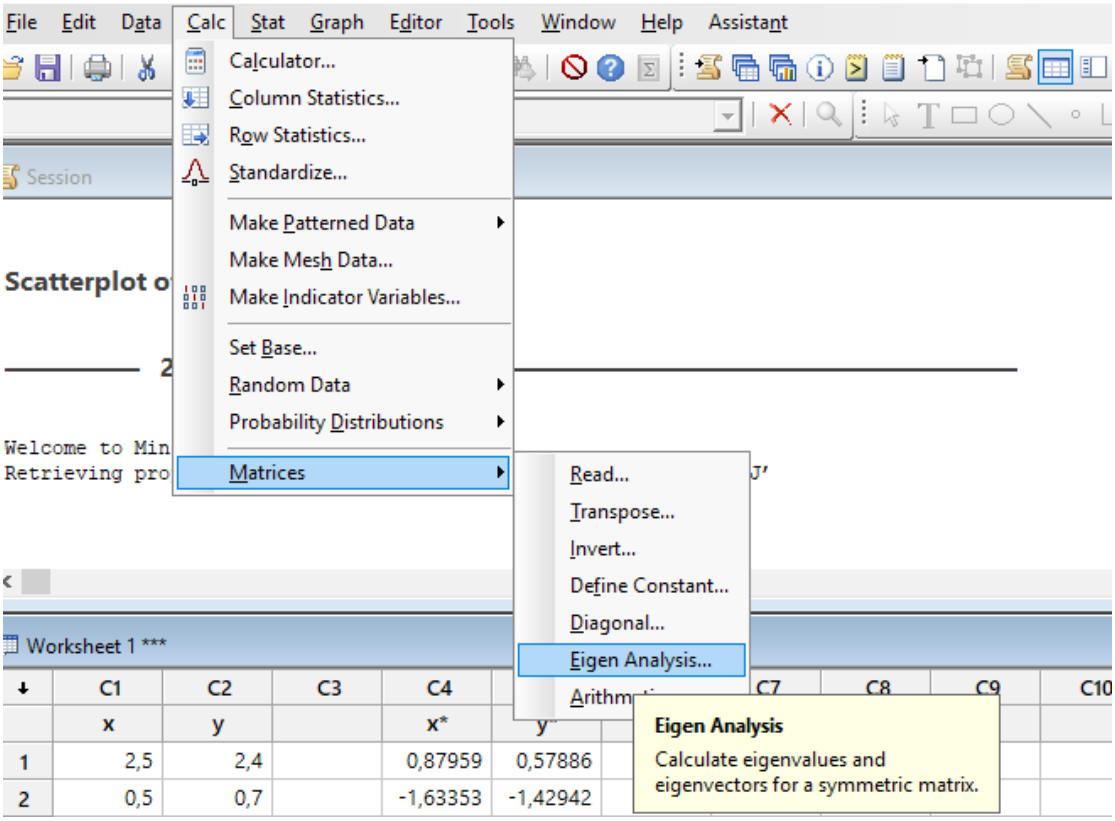
	x*	y*
x*	1,00000	
y*	0,93024	1,00000

**Το κενό κελί δηλώνει ότι είναι ίσο με το συμμετρικό του.

ΒΗΜΑ 3 Υπολογισμός των ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα συσχέτισης

Υπολογίζουμε τις ιδιοτιμές και τα ιδιοδιανύσματα του C και διατάσουμε τα αντίστοιχα ζευγάρια από τη μεγαλύτερη στη μικρότερη ιδιοτιμή.

minitab: **Calc ->matrices-> eigen analysis** και **επιλέγω τον πίνακα** C

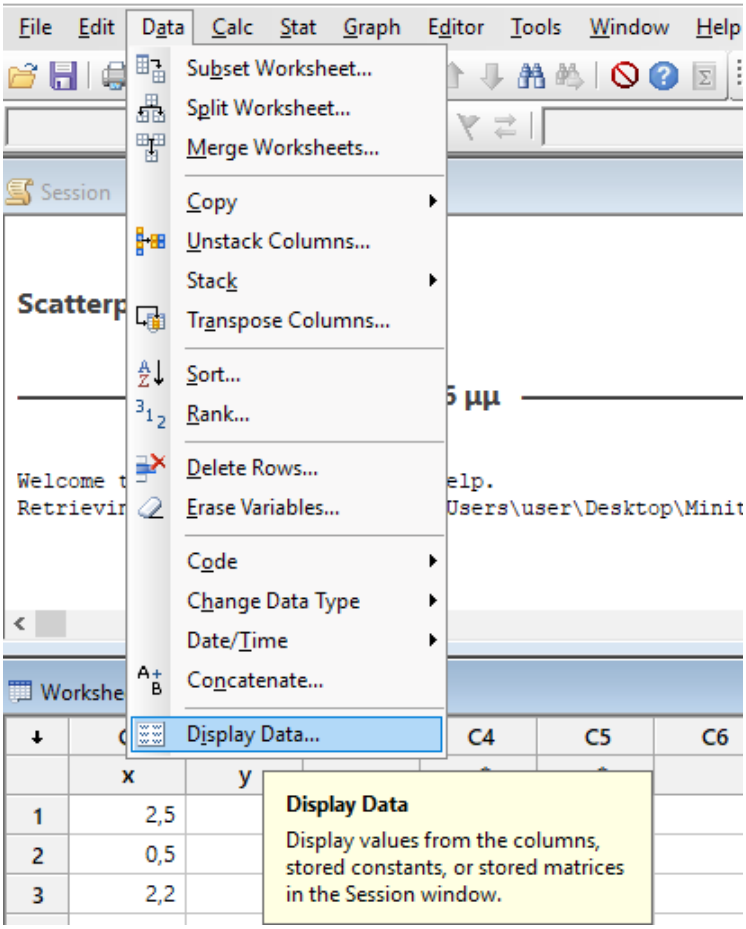


Οι ιδιοτιμές που παίρνουμε είναι:

$$\lambda_1 = \mathbf{1,2708} \quad \textbf{και} \quad \lambda_2 = \mathbf{0,0525}$$

και τα αντίστοιχα ιδιοδιανύσματα δίνονται με τη βοήθεια των εντολών:

Data → Display Data → επιλέγω όνομα πίνακα



και παίρνουμε τα ιδιοδιανύσματα:

$$\upsilon_1 = (0,674,\ 0,738)$$
$$\upsilon_2 = (0,454,\ 0,546)$$

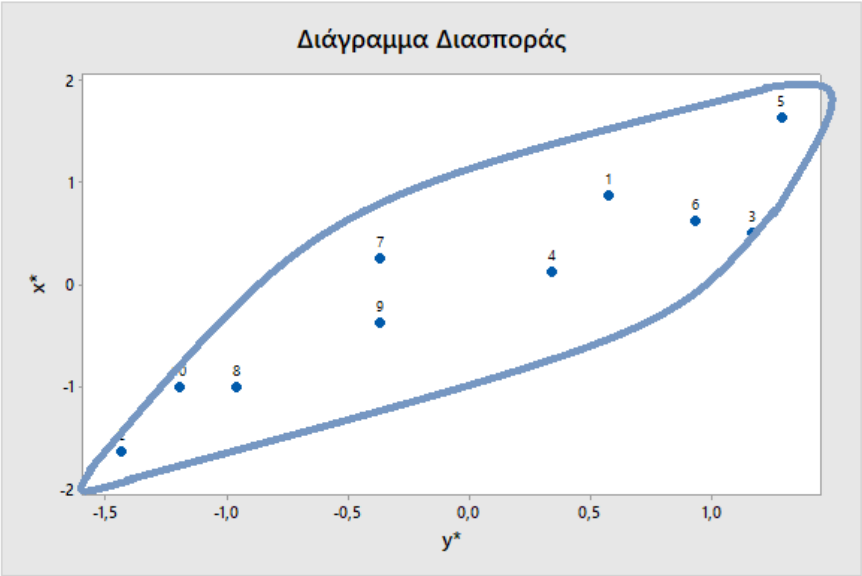
Παρατηρούμε ότι τα μήκη κάθε ενός διανύσματος είναι 1:

$$(\sqrt{0,674^2 + 0,738^2} = \sqrt{0,454^2 + 0,546^2} = 1).$$

Έχουμε λοιπόν:

	PC1	**PC2**
μεταβλητή	Ιδιοτιμή 1	Ιδιοτιμή 2
x	0,674	0,454
y	0,738	0,546
ιδιοτιμές	**1,2708**	**0,0525**
% **συνόλου διακύμανσης**	0,960	0,04

$$\frac{\lambda_1}{\lambda_1+\lambda_2}=\frac{1,2708}{1,2708+0,0525}=0,960 \text{ και } \frac{\lambda_2}{\lambda_1+\lambda_2}=\frac{0,525}{1,2708+0,0525}=0,04$$



Το πρώτο ιδιοδιάνυσμα θα μας δώσει τη διεύθυνση της σημαντικότερης συνιστώσας. Το δεύτερο ιδιοδιάνυσμα θα μας δώσει τη λιγότερο σημαντική συνιστώσα. Η μια συνιστώσα είναι ορθογώνια στην άλλη.

Ένα αρχικό συμπέρασμα που μπορούμε να βγάλουμε από το διάγραμμα είναι: η διεύθυνση ως προς την οποία αποκλίνει περισσότερο το νέφος από το σημείο προσδιορίζει την επικρατέστερη τάση απομάκρυνσης από την μέση κατάσταση παρατηρήσεων -- PC1

Ο άλλος άξονας δίνει λιγότερο έντονη τάση. --PC2

ΒΗΜΑ 4 Επιλογή αριθμού κυρίων συνιστωσών

Θα επιλέξουμε τώρα τον αριθμό των κυρίων συνιστωσών, χρησιμοποιώντας τις ιδιοτιμές κατά φθίνουσα τάξη μεγέθους. .

Ο άξονας που δημιουργείται από το πρώτο ιδιοδιάνυσμα ερμηνεύει το 96% της ολικής αδράνειας διακύμανσης των δεδομένων, ενώ ο δεύτερος ερμηνεύει μόνον το 4%.

Θα χρησιμοποιήσουμε λοιπόν το πρώτο ιδιοδιάνυσμα (αυτό που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή του πίνακα συσχετίσεων) σα διεύθυνση του πρώτου άξονα (κύρια συνιστώσα).

$$V=\begin{bmatrix}0.674\\0.738\end{bmatrix}$$

Βήμα 5 Προβολή δεδομένων

Εάν πολλαπλασιάσουμε (από δεξιά) τον πίνακα των αρχικών μας δεδομένων με τον πίνακα V προκύπτει ο πίνακας των συντεταγμένων των δεδομένων μας ως προς το νέο άξονα (κύρια συνιστώσα). Το αποτέλεσμα φαίνεται στον παρακάτω πίνακα.

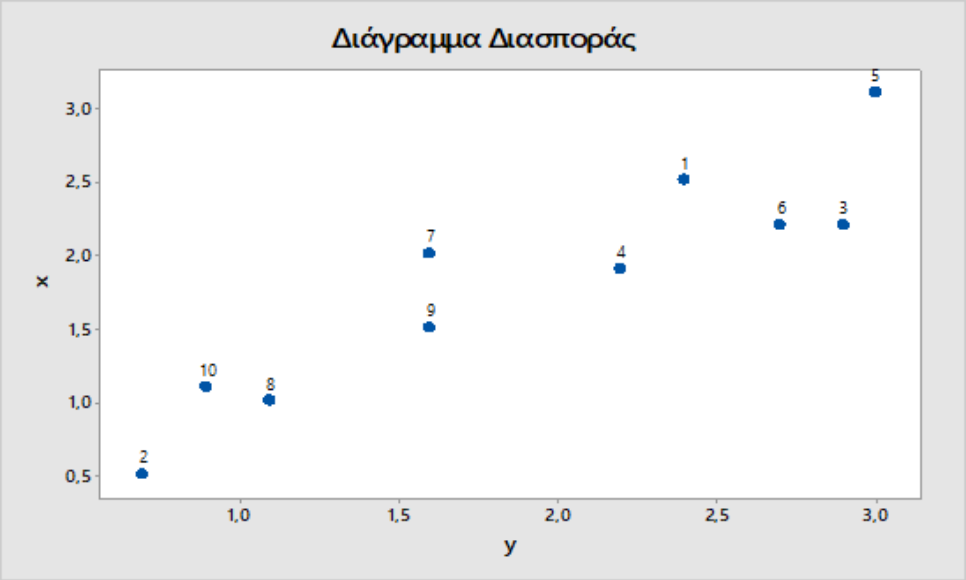
$$Y=XV=\begin{bmatrix}0.879593 & 0.578857\\ -1.63353 & -1.42942\\ 0.502625 & 1.16953\\ 0.125656 & 0.342589\\ 1.63353 & 1.28766\\ 0.628281 & 0.933259\\ 0.251312 & -0.366216\\ -1.00525 & -0.956886\\ -0.376968 & -0.366216\\ -1.00525 & -1.19315\end{bmatrix}\begin{bmatrix}0.674\\0.738\end{bmatrix}=\begin{bmatrix}3.4562\\0.8536\\3.623\\2.9042\\4.3034\\3.5428\\2.5288\\1.4858\\2.1918\\1.3382\end{bmatrix}$$

Έτσι, το σημείο 1 (2.5, 2.4), μετά τη μείωση των διαστάσεων σε μια συνιστώσα αντιστοιχεί στο σημείο 3,4562. Το σημείο 2 (0.5, 0.7) αντιστοιχεί στο σημείο 0,8536 κ.ο.κ. Γενικότερα, ένα σημείο με συντεταγμένες (x,y) στους «παλαιούς» άξονες, αντιστοιχεί στο σημείο:

Y=0,674x+0,738y

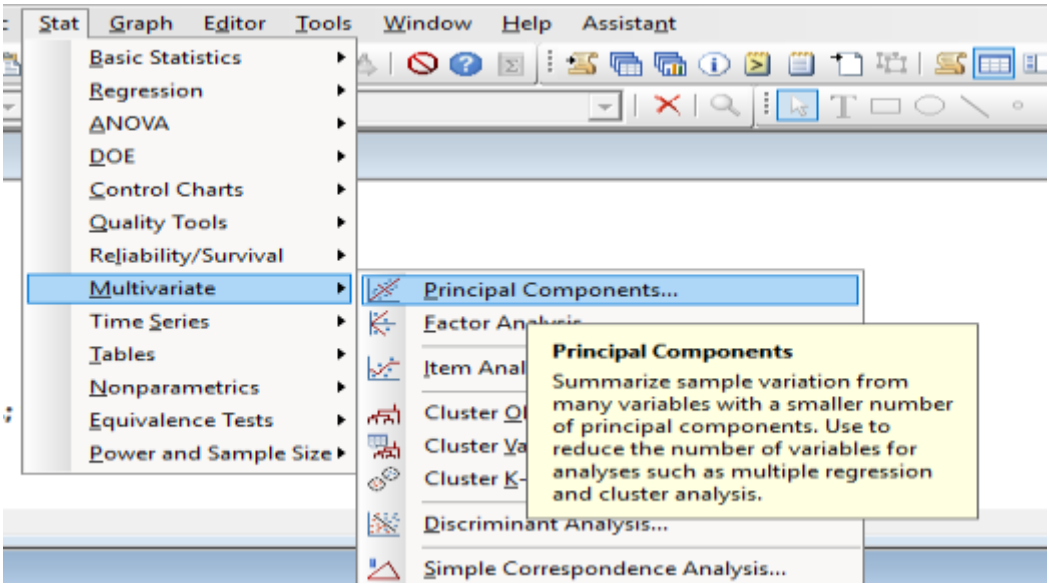
στο νέο άξονα

Αυτή η μετακίνηση μπορεί να διαπιστωθεί και από τα παρακάτω γράφημα:



Θα επαληθεύσουμε τις τιμές που βρήκαμε παραπάνω, χρησιμοποιώντας το minitab και ακολουθώντας τις παρακάτω εντολές:

STAT-> MULTIVARIATE->PRINCIPAL COMPONETS



Επιλέγοντας **number of componets**: 1 έχουμε

Principal Component Analysis: x; y

Eigenanalysis of the Covariance Matrix

Eigenvalue	1,2708	0,0525
Proportion	0,960	0,040
Cumulative	0,960	1,000

Variable	PC1
x	0,674
y	0,738

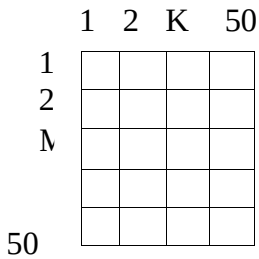
Κεφάλαιο 3: Εφαρμογή της Α.Κ.Σ. σε συστήματα αναγνώρισης προσώπου (facial recognition systems)

Η Ανάλυση Κυρίων Συνιστωσών μπορεί να χρησιμοποιηθεί για την αυτόματη αναγνώριση ή ταυτοποίηση ενός προσώπου από μια ψηφιακή εικόνα (facial recognition). Η αναγνώριση εφαρμόζεται συνήθως σε συστήματα ασφαλείας. Η ταυτοποίηση αυτή θα μπορούσε να πραγματοποιηθεί συγκρίνοντας χαρακτηριστικά της ψηφιακής εικόνας με τα χαρακτηριστικά άλλων εικόνων που είναι αποθηκευμένες σε μία βάση δεδομένων.

Ιστορικά, οι Sirovich και Kirby το 1987 χρησιμοποίησαν την Α.Κ.Σ. για να εξασφαλίσουν μια συμπιεσμένη αναπαράσταση ψηφιακών εικόνων προσώπου. Το 1991 οι Turk και Pentlad χρησιμοποίησαν την Α.Κ.Σ. (προβολές σε υποχώρους των ιδιοδιανυσμάτων – ιδιόχωρος ή eigenspace) για να λύσουν προβλήματα αναγνώρισης προσώπου χρησιμοποιώντας σα μέτρο ομοιότητας την Ευκλείδια απόσταση. Η μέθοδος τους ονομάστηκε **eigenfaces** και ήταν μια πρώτη προσπάθεια στην κατεύθυνση της αναγνώρισης προσώπων. Υπάρχουν διάφορες προσπάθειες επίλυσης του προβλήματος, οι οποίες ουσιαστικά διαφέρουν στο ότι χρησιμοποιούν διαφορετικούς τρόπους προβολής των εικόνων και διαφορετικές αποστάσεις ομοιότητας

3.1 Ιδέα των συστημάτων αναγνώρισης προσώπων

Μια οποιαδήποτε ασπρόμαυρη εικόνα μπορεί να θεωρηθεί σαν ένα διάνυσμα εικονοστοιχείων (pixels). Κάθε τιμή/συντεταγμένη του διανύσματος αντιστοιχεί σε ένα pixel και παριστάνει την τιμή κλίμακας του γκρι του αντίστοιχου pixel (σε μια κλίμακα από 0 έως 256). Για έγχρωμες εικόνες η διαδικασία είναι ανάλογη (αντιστοίχηση περισσότερων από μια τιμές σε κάθε pixel, για να περιγραφούν επίσης τα βασικά χρώματα). Έτσι π.χ. μια εικόνα είναι διάστασης 50x50 pixels μπορεί να παρασταθεί σαν ένα διάνυσμα 2500 στοιχείων/συντεταγμένων.

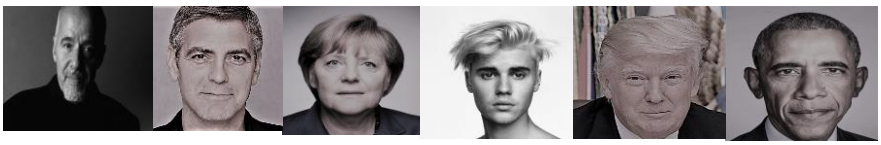


Εάν χρησιμοποιήσουμε έναν αριθμό αρχικών εικόνων (*εικόνες εκπαίδευσης*), έχουμε έναν αρχικό πίνακα δεδομένων (κάθε στήλη είναι η διανυσματική αναπαράσταση μιας εικόνας, όπως περιγράφηκε παραπάνω). Με τη βοήθεια της Α.Κ.Σ. μπορούμε να υπολογίσουμε τις κύριες συνιστώσες των δεδομένων μας. Εάν μας δοθεί μια νέα εικόνα/πρόσωπο (εικόνα ελέγχου) το αναπαριστάνουμε διανυσματικά με τον ίδιο τρόπο. Στη συνέχεια το προβάλουμε (βρίσκουμε δηλαδή τις συντεταγμένες) στο νέο σύστημα αξόνων/κυρίων συνιστωσών. Παίρνουμε την απόσταση της προβολής αυτής με την κάθε μια από τις προβολές των εικόνων εκπαίδευσης (σαν μέτρο απόστασης χρησιμοποιούμε συνήθως την ευκλείδια απόσταση). Η εικόνα εκπαίδευσης με τη μικρότερη απόσταση από την προβολή της εικόνας μας είναι (ίσως) η εικόνα του προσώπου που ζητάμε (στην πραγματικότητα η προβολή της).

Η παραπάνω διαδικασία, ανά βήμα, περιγράφεται αναλυτικά στη συνέχεια.

3.2 Αποθήκευση των εικόνων εκπαίδευσης

Έστω ότι έχουμε ένα σύνολο από M εικόνες (εικόνες εκπαίδευσης), κάθε μια διάστασης nxn (δηλαδή κάθε εικόνα αποτελείται από $n \cdot n = n^2$ pixels). Ένα παράδειγμα εικόνων εκπαίδευσης είναι και το παρακάτω.



Έτσι κάθε εικόνα μπορεί να παρασταθεί σαν ένα διάνυσμα διάστασης $p = n \cdot n = n^2$

Δηλαδή $\Gamma_i = \begin{pmatrix} a_{i1} \\ a_{i2} \\ \dots \\ a_{in^2} \end{pmatrix}$, $i = 1, 2, \dots, M$. Τα διανύσματα αυτά αποτελούν τις στήλες του

πίνακα δεδομένων μας A (διάστασης pxM).

$$A = \begin{pmatrix} a_{11} & \text{K} & a_{n^2 1} \\ \text{M} & \text{O} & \text{M} \\ a_{n^2 1} & \text{L} & a_{n^2 M} \end{pmatrix}$$

3.3 Κανονικοποίηση των διανυσμάτων εικόνας

Κανονικοποιούμε τον πίνακα αφαιρώντας από κάθε pixel τη μέση τιμή των τιμών που αντιστοιχούν στον ίδιο αύξοντα αριθμό του pixel σε όλες τις εικόνες (μ' άλλα λόγια θα αφαιρέσουμε από κάθε εικόνα όλα τα κοινά χαρακτηριστικά που έχουν οι εικόνες αυτές και θα μείνουν σε κάθε εικόνα μόνο τα διαφορετικά τους χαρακτηριστικά , δηλαδή μόνο τα ιδιαίτερα τους χαρακτηριστικά)

Το μέσο διάνυσμα (κοινά χαρακτηριστικά των εικόνων) υπολογίζεται από τη σχέση:

$$\Psi = \frac{1}{M}\sum_{i=1}^{M}\Gamma_i$$

Στη συνέχεια: **κανονικοποιήμενο διάνυσμα = διάνυσμα εικόνας – μέσο διάνυσμα**

$$\Phi_i = \Gamma_i - \Psi$$

Δημιουργούμε έτσι τον πίνακα A_0 (των κεντρικοποιημένων εικόνων). Θέτουμε:

$$X = \frac{1}{M}A_0 \quad \text{ή} \quad X = A_0$$

3.4 Εύρεση του πίνακα διακύμανσης των εικόνων εκπαίδευσης

Υπολογίζουμε τον πίνακα (συν) διακύμανσης των δεδομένων, που είναι ίσος με:

$$C = XX^T$$

(ο πίνακας X είναι διάστασης $n^2 xM$ και ο X^T διάστασης Mxn^2, οπότε ο $C = XX^T$ είναι διάστασης $n^2 xn^2$).

Στο παράδειγμα μας ο πίνακας (συν)διακύμανσης θα είναι διάστασης 2500x2500 άρα θα έχουμε έναν πίνακα 2500 γραμμών και 2500 στηλών που σημαίνει ότι θα έχουμε (μέχρι) 2500 ιδιοδιανύσματα να υπολογίσουμε όπου το καθένα θα έχει διάσταση 2500x1. Γενικότερα, θα έχουμε να υπολογίσουμε (μέχρι) $p = n^2$ ιδιοδιανύσματα.

Παρατήρηση 3.4.1

Θυμόμαστε ότι, στην Α.Κ.Σ., θα πρέπει να επιλέξουμε τα k «σημαντικότερα» ιδιοδιανύσματα (eigenfaces) από το σύνολο των ιδιοδιανυσμάτων όπου το k θα πρέπει να είναι μικρότερο ή ίσο του $p = n^2$.

Δηλαδή στην περίπτωση που θα έχουμε 10 φωτογραφίες όπου η κάθε μία είναι διάστασης 50x50 εμείς θα έχουμε να διαλέξουμε τα σημαντικότερα από 2500 ιδιοδιανύσματα. Η διαδικασία αυτή μπορεί να επιβραδύνει και να εξαντλήσει ακόμα τη μνήμη του μέσου στο οποίο εργαζόμαστε. Οι υπολογισμοί που χρειάζονται είναι τεράστιοι! Η λύση του προβλήματος: μείωση διαστάσεων

Για να επιτευχθεί λοιπόν αυτό απλά θα χρησιμοποιήσουμε τον πίνακα:

$$\dot{C} = X^T X$$

(στο παράδειγμα με τις 10 εικόνες, ο X^T είναι διάστασης (10x2500) και ο X διάστασης (2500x10), άρα ο πίνακας (συν)διακύμανσης $\dot{C} = X^T X$ θα είναι 10x10. Σε αυτή την περίπτωση το σύνολο των ιδιοδιανυσμάτων θα είναι (το πολύ) 10, όσες και οι διαφορετικές εικόνες. Είναι λοιπόν τώρα πολύ πιο εύκολο να διαλέξουμε k ιδιοδιανύσματα από τον αριθμό των 10. Δηλαδή η διάσταση των ιδιοδιανυσμάτων μειώθηκε από (2500x1) σε (10x1).

Παρατήρηση 3.4.2

Είναι γνωστό από τη γραμμική άλγεβρα ότι οι πίνακες $C = XX^T$ και $\dot{C} = X^T X$ έχουν τις ίδιες ιδιοτιμές. Επιπλέον τα ιδιοδιανύσματα του πίνακα $C = XX^T$ προκύπτουν από τα ιδιαδιανύσματα του πίνακα $\dot{C} = X^T X$ αφού τα πολλαπλασιάσουμε με τον πίνακα X και τα κανονικοποιήσουμε (τα διαιρέσουμε με το μέτρο τους). Αν δηλαδή V' είναι ο πίνακας που έχει σαν στήλες τα ιδιοδιανύσματα του πίνακα $\dot{C} = X^T X$, τότε τα ιδιοδιανύσματα του πίνακα $C = XX^T$ είναι οι στήλες του πίνακα XV' (αφού διαιρέσουμε κάθε διάνυσμα με το μέτρο του).

3.5 Επιλογή k «καλύτερων» ιδιοδιανύσματων του πίνακα συνδιακύμανσης.

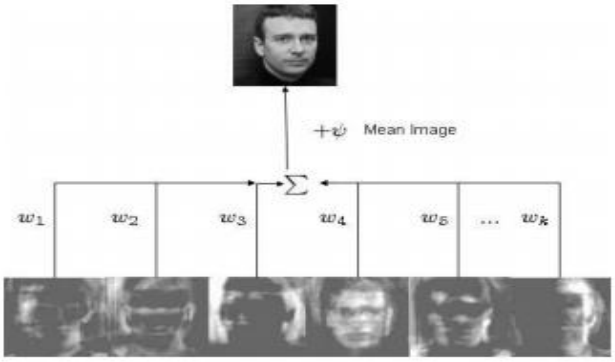
Επιλέγουμε στη συνέχεια τα k «καλύτερα» ιδιοδιανύσματα του πίνακα (συν) διακύμανσης $\dot{C} = X^T X$, δηλαδή τα διανύσματα αυτά που θα αντιστοιχούν σε μη μηδενικές ιδιοτιμές και θα ερμηνεύουν μεγάλο ποσοστό της διακύμανσης των αρχικών εικόνων. Με αυτά θα δημιουργήσουμε το λεγόμενο **πίνακα προβολής**,

οποίος θα ορίζει και τον ιδιοχώρο των εικόνων (υπόχωρος του αρχικού χώρου, μικρότερης διάστασης). Έστω ότι τον ονομάζουμε πίνακα $\dot{V}'$.

Εάν πολλαπλασιάσουμε τον πίνακα $\dot{V}'$ με τον X, παίρνουμε ένα πίνακα $\dot{V}$ (δηλαδή $\dot{V} = \left[\dot{v}_1 \ \dot{v}_2 \ldots \dot{v}_k \right]$) που είναι οι συντεταγμένες των κύριων συνιστωσών εάν χρησιμοποιήσουμε σαν πίνακα διακύμανσης τον $C = XX^T$.

Κάθε εικόνα εκπαίδευσης Γ_i, $i = 1, 2, ..., M$ προβάλλεται στον ιδιόχωρο των κύριων συνιστωσών πολλαπλασιάζοντας τον πίνακα $\left(\dot{V} \right)^T$ με το Γ_i. Συμβολίζουμε με

$$\Omega_i = \left(\dot{V} \right)^T \cdot \Gamma_i = \begin{pmatrix} \omega_1 \\ \omega_1 \\ \ldots \\ \omega_k \end{pmatrix},$$ (διάνυσμα βάρους της κάθε εικόνας εκπαίδευσης).



3.6 Αναγνώριση εικόνας ελέγχου

Έστω ότι μας δίνεται μια *εικόνα ελέγχου*. Προκειμένου να γίνει η αναγνώριση της (ταυτοποίηση ίσως με κάποια εικόνα εκπαίδευσης), η εικόνα ελέγχου θεωρείται σαν ένα διάνυσμα όπως πριν. Κανονικοποιούμε την εικόνα αφαιρώντας της το μέσο πρόσωπο. Στη συνέχεια προβάλουμε την εικόνα στον υπόχωρο των κύριων συνιστωσών.

Αν λοιπόν Υ το κανονικοποιημένο διάνυσμα της εικόνας ελέγχου, η προβολή της (το διάνυσμα βάρους της) θα είναι

$$\tilde{Y} = \left(\vdots V\right)^T Y$$

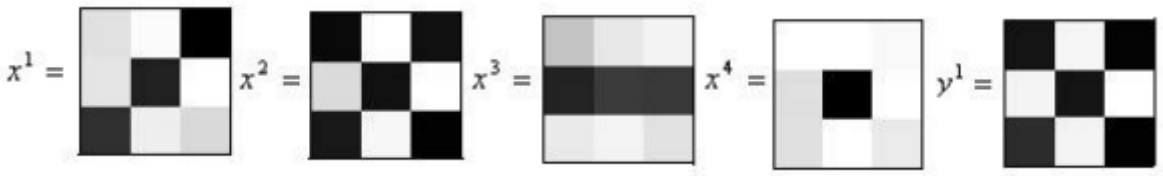
Στην συνέχεια υπολογίζουμε την απόσταση του διάνυσμα βάρους $\tilde{Y}$ από το διάνυσμα βάρος Ω_i της κάθε εικόνας εκπαίδευσης. Για τον υπολογισμό της απόστασης χρησιμοποιήσουμε συνήθως την ευκλείδεια απόσταση.

$$d(\Omega_i, \tilde{Y}) = \sqrt{\left(\omega_1 - \vdots y_1\right)^2 + \left(\omega_2 - \vdots y_2\right)^2 + \ldots + \left(\omega_k - \vdots y_k\right)^2}$$

Από τις διαφορές επιλέγεται η ελάχιστη. Δηλαδή η εικόνα εκπαίδευσης που αντιστοιχεί στην ελάχιστη διαφορά που βρέθηκε είναι παρόμοια ή πιο κοντά στην εικόνα ελέγχου.

Παράδειγμα 3.6.1

Έστω ότι έχουμε τις παρακάτω 4 εικόνες από τις οποίες η κάθε μια είναι διάστασης 3x3 pixel, καθώς και την τελευταία η οποία είναι η εικόνα ελέγχου.



Βήμα 1

Κάθε μια από τις εικόνες παριστάνεται σαν ένα διάνυσμα 9 διαστάσεων (η κάθε συντεταγμένη του διανύσματος είναι μια τιμή που παριστάνει τη διαβάθμιση του γκρι του αντίστοιχου pixel της εικόνας). Οι 4 εικόνες αποτελούν τις στήλες του παρακάτω πίνακα.

$$A = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ 225 & 10 & 196 & 255 \\ 229 & 219 & 35 & 223 \\ 48 & 24 & 234 & 224 \\ 251 & 255 & 232 & 255 \\ 33 & 18 & 59 & 0 \\ 238 & 247 & 244 & 255 \\ 0 & 17 & 247 & 249 \\ 255 & 255 & 57 & 255 \\ 217 & 2 & 226 & 235 \end{bmatrix}$$

Βήμα 2

Κανονικοποιούμε τα διανύσματα x_1, x_2, x_3, x_4 . Θα αφαιρέσουμε από το καθένα το μέσο διάνυσμα

$$\begin{bmatrix} \mu \\ 171,5 \\ 176,5 \\ 132,5 \\ 248,25 \\ 27,5 \\ 246 \\ 128,25 \\ 205,5 \\ 170 \\ \end{bmatrix}$$

και έτσι θα έχουμε τις κανονικοποιημένες εικόνες- διανύσματα. $\overset{g}{x_i} = x_i - \mu$

$$X = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ 53.5 & -161.5 & 24.5 & 83.5 \\ 52.5 & 42.5 & -141.5 & 46.5 \\ -84.5 & -108.5 & 101.5 & 91.5 \\ 2.75 & 6.75 & -16.25 & 6.75 \\ 5.5 & -9.5 & 31.5 & -27.5 \\ -8 & 1 & -2 & 9 \\ -128.25 & -111.25 & 118.75 & 120.75 \\ 49.5 & 49.5 & -148.5 & 49.5 \\ 47 & -168 & 56 & 65 \end{bmatrix}$$

Συνδυάζουμε όλες τις κανονικοποιημένες τιμές σε έναν πίνακα που τον ονομάζουμε X.

Βήμα 3

Στη συνέχεια υπολογίζουμε τον πίνακα συνδιασποράς του παραπάνω πίνακα:

$$C = X^T X$$

Ο πίνακας X^T είναι διάστασης 4x9, ενώ ο πίνακας X είναι διάστασης 9x4. Επομένως ο πίνακας $C = X^T X$ θα είναι διάστασης 4x4 .

$$\mathbf{C}= \begin{bmatrix} 53.5 & 52.5 & -84.5 & 2.75 & 5.5 & -8 & -128.25 & 49.5 & 47 \\ -161.5 & 42.5 & -108.5 & 6.75 & -9.5 & 1 & -111.25 & 49.5 & -168 \\ 24.5 & -141.5 & 101.5 & -16.25 & 31.5 & -2 & 115.75 & -148.5 & 56 \\ 83.5 & 46.5 & 91.5 & 6.75 & -27.5 & 9 & 121.75 & 49.5 & 65 \end{bmatrix} \bullet$$

$$\begin{bmatrix} 53.5 & -161.5 & 24.5 & 83.5 \\ 52.5 & 42.5 & -141.5 & 46.5 \\ -84.5 & -108.5 & 101.5 & 91.5 \\ 2.75 & 6.75 & -16.25 & 6.75 \\ 5.5 & -9.5 & 31.5 & -27.5 \\ -8 & 1 & -2 & 9 \\ -128.25 & -111.25 & 118.75 & 120.75 \\ 49.5 & 49.5 & -148.5 & 49.5 \\ 47 & -168 & 56 & 65 \end{bmatrix} =$$

$$\begin{bmatrix} 33.967,875 & 11539,625 & -34498,625 & -11008,875 \\ 11539,625 & 82848,375 & -51363,875 & -43024,125 \\ -34498,625 & -51323,875 & 71474,875 & 14387,625 \\ -11008,875 & -43024,125 & 14387,625 & 39645,375 \end{bmatrix}$$

Οι ιδιοτιμές του πίνακα είναι:

$$\lambda_1 = 3,3364 \quad \lambda_2 = 0,5510, \ \lambda_3 = 0,1126 \quad \lambda_4 = 0,0000$$

Βήμα 4

Θα υπολογίσουμε τα ιδιοδιανύσματα που αντιστοιχούν σε μη μηδενικές ιδιοτιμές του πίνακα συσχέτισης. Έστω $v_i, i=1,2,3$ το ιδιοδιάνυσμα που αντιστοιχεί στην ιδιοτιμή $\lambda_i, i=1,2,3$. Οι συντεταγμένες καθενός από αυτά τα διανύσματα αποτελούν τις στήλες του παρακάτω πίνακα V. Ο υπόχωρος που ορίζουν αυτά τα διανύσματα καλείται ιδιόχωρος των εικόνων εκπαίδευσης.

$$V = \begin{bmatrix} -0.468321882 & -0.66036106 & -0.498288332 \\ -0.521298424 & 0.3676719131 & 0.4090553232 \\ 0.5235254636 & 0.3095014953 & -0.539635843 \\ 0.4846067031 & -0.577017729 & 0.5414569577 \end{bmatrix}$$

Βήμα 5

Θα υπολογίσουμε τώρα τα (κυριότερα) ιδιοδιανύσματα του πίνακα $\dot{C} = X^T X$. Αυτό θα επιτευχθεί με τον μετασχηματισμό

$$U = A \bullet V = \begin{bmatrix} 115,600 & -231,382 & -75,721 \\ -95,019 & -188,545 & 77,333 \\ 196,066 & -79,702 & -19,089 \\ -5,447 & -147,329 & -7,885 \\ 6,050 & 3,087 & -40,919 \\ 11,094 & -137,972 & -11,156 \\ 241,116 & -60,980 & 8,487 \\ -98,938 & -204,134 & 84,558 \\ 129,531 & -208,215 & -102,026 \end{bmatrix}$$

Βήμα 6

Για κάθε εικόνα θα υπολογίσουμε το διάνυσμα βάρους της.

Για την εικόνα 1: $\Omega_1 = U^T \overset{g}{X}_1 = \begin{bmatrix} -45174.5 \\ -26897.1 \\ -233.6 \end{bmatrix}$

Για την εικόνα 2: $\Omega_2 = U^T \overset{g}{X}_2 = \begin{bmatrix} -97546.8 \\ 68500.4 \\ 38293.0 \end{bmatrix}$

Για την εικόνα 3: $\Omega_3 = U^T \overset{g}{X}_3 = \begin{bmatrix} 87013,5 \\ 27100,2 \\ -33136,3 \end{bmatrix}$

Για την εικόνα 4: $\Omega_4 = U^T \overset{g}{X}_4 = \begin{bmatrix} 55707,8 \\ -68703,5 \\ -4923,1 \end{bmatrix}$

Βήμα 7

Η εικόνα ελέγχου που αντιμετωπίζεται σαν διάνυσμα πριν και μετά την κανονικοποίηση (αφαιρώντας της την μέση εικόνα) είναι:

Y	$\overset{g}{Y}$
20	−151.5
244	67.5
44	−88.5
246	−2.25
21	−6.5
244	−2
4	−124.25
255	49.5
2	−168

Το διάνυσμα βάρους της εικόνας ελέγχου είναι:

$$\tilde{Y} = U^T Y = \begin{bmatrix} -97945,6 \\ 62420,8 \\ 38958,6 \end{bmatrix}$$

Συγκρίνουμε την απόσταση του από κάθε μία από τις εικόνες:

$$d(\Omega_1, \tilde{Y}) = \sqrt{((-45174,5)-(-979945,6))^2 + (-26897,1-62420,8)^2 + (-233,6-38958,6)^2}$$

$$=110898,66544$$

Συνεχίζοντας, με τον ίδιο τρόπο βρίσκουμε και τις υπόλοιπες αποστάσεις.

$$d(\Omega_2, \tilde{Y}) = 6128,915208$$

$$d(\Omega_3, \tilde{Y}) = 201631,09647$$

$$d(\Omega_4, \tilde{Y}) = 206708,84607$$

Παρατηρούμε ότι η ελάχιστη απόσταση της $\tilde{Y}$ είναι από την εικόνα 2. Φτάνουμε λοιπόν στο συμπέρασμα ότι η εικόνα ελέγχου ταυτοποίησης είναι παρόμοια ή **παρουσιάζει πολλά κοινά με την εικόνα 2**

Κεφάλαιο 4: Εφαρμογή της Ανάλυσης Κύριων Συνιστωσών σε δεδομένα από τα Ελληνικά Α.Ε.Ι

Παράδειγμα 4.1

Για κάθε ένα από τα Ανώτατα Εκπαιδευτικά Ιδρύματα της χώρας (παρατήρηση) έχουμε δεδομένα για 8 μεταβλητές. που αφορούν στοιχεία τους.

Μεταβλητή	Περιγραφή
Μέλη ΔΕΠ	Αριθμός διδακτικού ερευνητικού προσωπικού
Προπτυχιακοί	Αριθμός προπτυχιακών φοιτητών
Μεταπτυχιακοί	Αριθμός μεταπτυχιακών φοιτητών
Διδακτορικοί	Αριθμός διδακτορικών φοιτητών
Τμήματα	Αριθμός τμημάτων ανά εκπαιδευτικό ίδρυμα
Εισακτέοι	Αριθμός εισακτέων ανά εκπαιδευτικό ίδρυμα
Πόλεις	Αριθμός πόλεων διασποράς του ιδρύματος
Έτος ίδρυσης	Έτος ίδρυσης του ιδρύματος

Εισάγοντας τα παραπάνω δεδομένα στο minitab έχουμε τον παρακάτω πίνακα (του νέφους) 20 σημείων στο χώρο R^8.

		Μέλη ΔΕΠ	Προπτυχιακοί	Μεταπτυχιακοί	Διδακτορικοί	τμήματα	εισακτέοι	πόλεις	ετος ιδρυσης
1	πανεπ.Αθηνών	1764	29309	11792	4329	33	5455	1	1837
2	πανεπ.Αιγαίου	316	9086	1576	574	17	2855	6	1984
3	πανεπ.Θεσσαλίας	435	9702	1772	1137	18	2510	5	1984
4	πολυτεχν.Κρήτης	121	3219	479	260	5	720	1	1984
5	πανεπ.Θεσσαλονίκης	1944	32700	5466	3849	41	6000	1	1925
6	πανεπ.Θράκης	579	14394	2617	1802	19	3835	4	1973
7	πανεπ.Ιόνιο	111	2820	439	439	6	825	1	1984
8	πανεπ.Ιωαννίνων	510	11327	1519	1775	15	2930	1	1970
9	πανεπ.Κρήτης	485	11274	1351	1137	16	2710	2	1973
10	πανεπ.Πατρών	705	18411	1979	1795	24	4345	2	1964
11	Οικον.Πανεπ.Αθήνας	187	6813	1818	253	8	1465	1	1920
12	Πάντειο Πανεπ.	230	7490	1293	1456	9	1530	1	1927
13	Πανεπ.Πειραιώς	181	7995	3293	345	9	1745	1	1938
14	Πανεπ.Μακεδονίας	205	6771	1752	402	8	1480	1	1948
15	Γεωπον.Πανεπ.Αθηνών	177	2991	321	306	6	730	1	1920
16	Ανώτ.σχολή Καλών Τεχνών	40	842	150	0	2	80	1	1930
17	Χαροκόπειο Πανεπ.	67	1461	784	192	4	320	1	1990
18	Πανεπ.Πελοποννησου	139	4328	1668	415	9	1500	5	2000
19	Πανεπ.Δυτικής Μακεδονίας	73	2415	829	159	6	680	2	2002
20	Μετσόβιο Πολυτεχνείο	514	7614	1786	2923	9	1230	1	1836

Όπως διαπιστώνουμε από μια πρώτη ανάγνωση του πίνακα των δεδομένων, οι μεταβλητές είναι ετερογενείς ως προς τις μεταβολές τους (διαφορετικές μονάδες μέτρησης και διαφορετικές διασπορές), πράγμα μη-επιθυμητό. Με την ανάλυση σε

κύριες συνιστώσες το πρόβλημα επιλύεται εύκολα αν κανονικοποιήσουμε την κάθε μεταβλητή. Αντικαθιστούμε λοιπόν την τιμή x της κάθε μιάς μεταβλητής του πίνακα με την τιμή $x' = \frac{x - \bar{X}}{\sigma_X}$, δηλαδή αφαιρούμε τη μέση τιμή της μεταβλητής και διαιρούμε με την τυπική της απόκλιση.

Με τη βοήθεια του excel υπολογίζουμε τη μέση τιμή και την τυπική απόκλιση κάθε μεταβλητής.

(i) εντολή για τη μέση τιμή: **AVERAGE**

(ii) εντολή για τυπική απόκλιση **: STDEV**

και παίρνουμε τους παρακάτω πίνακες:

	Μέλη ΔΕΠ	Προπτυχιακ	Μεταπτυχ	Διδακτορι	τμήματα	εισακτέοι	πόλεις	ετος ιδρυσης
ΜΕΣΕΣ ΤΙΜΕΣ	439,15	9548,1	2134,2	1177,4	13,2	2147,25	1,95	1949,45
ΤΥΠ.ΑΠΟΚΛΙΣΗ	521,451948	8608,893314	2565,61909	1255,7515	10,0556347	1674,86565	1,63755273	47,2077993

	ΚΑΝΟΝΙΚΟΠΟΙΗΜΕΝΕΣ ΤΙΜΕΣ							
	Μέλη ΔΕΠ	Προπτυχιακ	Μεταπτυχ	Διδακτορι	τμήματα	εισακτέοι	πόλεις	ετος ιδρυσης
πανεπ.Αθηνών	2,540694315	2,2954053777	3,764315614	2,50973222	1,969045276	1,974934527	-0,580133991	-2,36083871
πανεπ.Αιγαίου	-0,236167494	-0,0536770504	-0,21756932	-0,480509082	0,377897578	0,422571207	2,473202802	0,753053532
πανεπ.Θεσσαλίας	-0,007958547	0,0178768623	-0,141174503	-0,03217197	0,477344309	0,216584536	1,862535443	0,753053532
πολυτεχν.Κρήτης	-0,610123332	-0,7351816049	-0,645146431	-0,730558554	-0,815463195	-0,852157903	-0,580133991	0,753053532
πανεπ.Θεσσαλονί	2,885884319	2,6893003742	1,298633929	2,127490988	2,764619124	2,300333761	-0,580133991	-0,496739953
πανεπ.Θράκης	0,268193456	0,5628946516	0,188180701	0,497391403	0,57679104	1,007692766	1,251868085	0,520041187
πανεπ.Ιόνιο	-0,629300554	-0,7815290256	-0,66073721	-0,588014428	-0,716016464	-0,789466307	-0,580133991	0,753053532
πανεπ.ιωαννίνων	0,135870621	0,2066351546	-0,23978618	0,475890333	0,179004116	0,467350919	-0,580133991	0,456492366
πανεπ.Κρήτης	0,087927565	0,2004787303	-0,305267451	-0,03217197	0,278450847	0,335997099	0,030533368	0,520041187
πανεπ.Πατρών	0,509826459	1,0295051502	-0,060492222	0,491817051	1,074024696	1,312194802	0,030533368	0,329394724
Οικον.Πανεπ.Αθήν	-0,483553664	-0,3177063417	-0,123245107	-0,736132905	-0,517123002	-0,407346106	-0,580133991	-0,602654655
Πάντειο Πανεπ.	-0,401091607	-0,2390667332	-0,32787408	0,221859182	-0,417676271	-0,368537023	-0,580133991	-0,454374072
Πανεπ.Πειραιώς	-0,495059997	-0,1804064639	0,451664865	-0,662870003	-0,417676271	-0,240168517	-0,580133991	-0,221361727
Πανεπ.Μακεδονίας	-0,449034663	-0,3225850176	-0,148969892	-0,617478856	-0,517123002	-0,398390163	-0,580133991	-0,009532323
Γεωπον.Πανεπ.Αθ	-0,502730886	-0,7616658453	-0,706730008	-0,693927103	-0,716016464	-0,846187275	-0,580133991	-0,602654655
Ανώτ.σχολή Καλών	-0,765458834	-1,011291427	-0,773380588	-0,937605888	-1,113803388	-1,234278105	-0,580133991	-0,390825251
Χαροκόπειο Πανετ	-0,713680333	-0,9393890375	-0,526266742	-0,784709395	-0,914909926	-1,090983029	-0,580133991	0,880151175
Πανεπ.Πελοττπον	-0,575604332	-0,6063613303	-0,181710528	-0,60712649	-0,417676271	-0,386448907	1,862535443	1,091980579
Πανεπ.Δυτικής Μα	-0,702174	-0,8285734	-0,508727116	-0,81098848	-0,716016464	-0,876040416	0,030533368	1,13434646
Μετσόβιο Πολυτεχν	0,14354151	-0,2246630235	-0,13571773	1,390083946	-0,417676271	-0,547655867	-0,580133991	-2,38202165

Ο τετραγωνικός *πίνακας της (αδράνειας) διακύμανσης* είναι ίσος με $V = X^T X$ δηλαδή ίσος με το γινόμενο του ανάστροφου του κανονικοποιημένου πίνακα επί τον κανονικοποιημένο (όπου X ο κανονικοποιημένος πίνακας των δεδομένων).

Εντολές excel:

(i) εύρεση ανάστροφου πίνακα: **TRANSPOSE**

(ii) εύρεση γινομένου 2 πινάκων: **MMULT**

	πίνακας διασποράς							
	19,00000001	18,4770690789	15,93341111	17,50907685	17,99253452	17,15405411	-1,71898501	-9,339200386
	18,47706908	19,0000000184	15,52686476	16,82366112	18,6030833	18,4164803	-0,325014688	-7,789975616
V=	15,93341111	15,5268647622	19,00000003	14,9120592	14,20238062	14,1049318	-1,642286907	-11,03840294
	17,50907685	16,8236611221	14,9120592	19,00000002	15,93160467	15,49307664	-2,332082728	-12,04443938
	17,99253452	18,6030833014	14,20238062	15,93160467	19,00000005	18,60640245	2,744945041	-5,263907561
	17,15405411	18,4164803041	14,1049318	15,49307664	18,60640245	18,99999992	3,192952656	-4,824915501
	-1,71898501	-0,3250146876	-1,642286907	-2,332082728	2,744945041	3,192952656	19,00000003	8,724503282
	-9,339200386	-7,7899756156	-11,03840294	-12,04443938	-5,263907561	-4,824915501	8,724503282	19,00897435

Ο πίνακας V αντιστοιχεί στις συνδιακυμάνσεις / συνδιασπορές των μεταβλητών, της ανάλυσης των 20 παρατηρήσεων στο χώρο των 8 μεταβλητών**.** Ο τετραγωνικός *πίνακας των συσχετίσεων* αυτών των μεταβλητών θα είναι ίσος με:

$$R = \frac{1}{n-1} X^T X$$

	πίνακας συσχέτισης							
	1,000000001	0,9724773199	0,796670555	0,92153036	0,946975501	0,902844953	-0,090472895	-0,491536862
	0,97247732	1,000000001	0,776343238	0,885455849	0,979109647	0,969288437	-0,017106036	-0,409998717
	0,838600585	0,8172034085	0,950000002	0,784845221	0,747493717	0,742364832	-0,086436153	-0,580968576
R=	0,92153036	0,8854558485	0,74560296	1,000000001	0,838505509	0,815425086	-0,122741196	-0,633917862
	0,946975501	0,9791096474	0,710119031	0,838505509	1,000000003	0,97928434	0,144470792	-0,277047766
	0,902844953	0,9692884371	0,70524659	0,815425086	0,97928434	0,999999996	0,16805014	-0,253942921
	-0,090472895	-0,0171060362	-0,082114345	-0,122741196	0,144470792	0,16805014	1,000000001	0,459184383
	-0,491536862	-0,4099987166	-0,551920147	-0,633917862	-0,277047766	-0,253942921	0,459184383	1,000472334

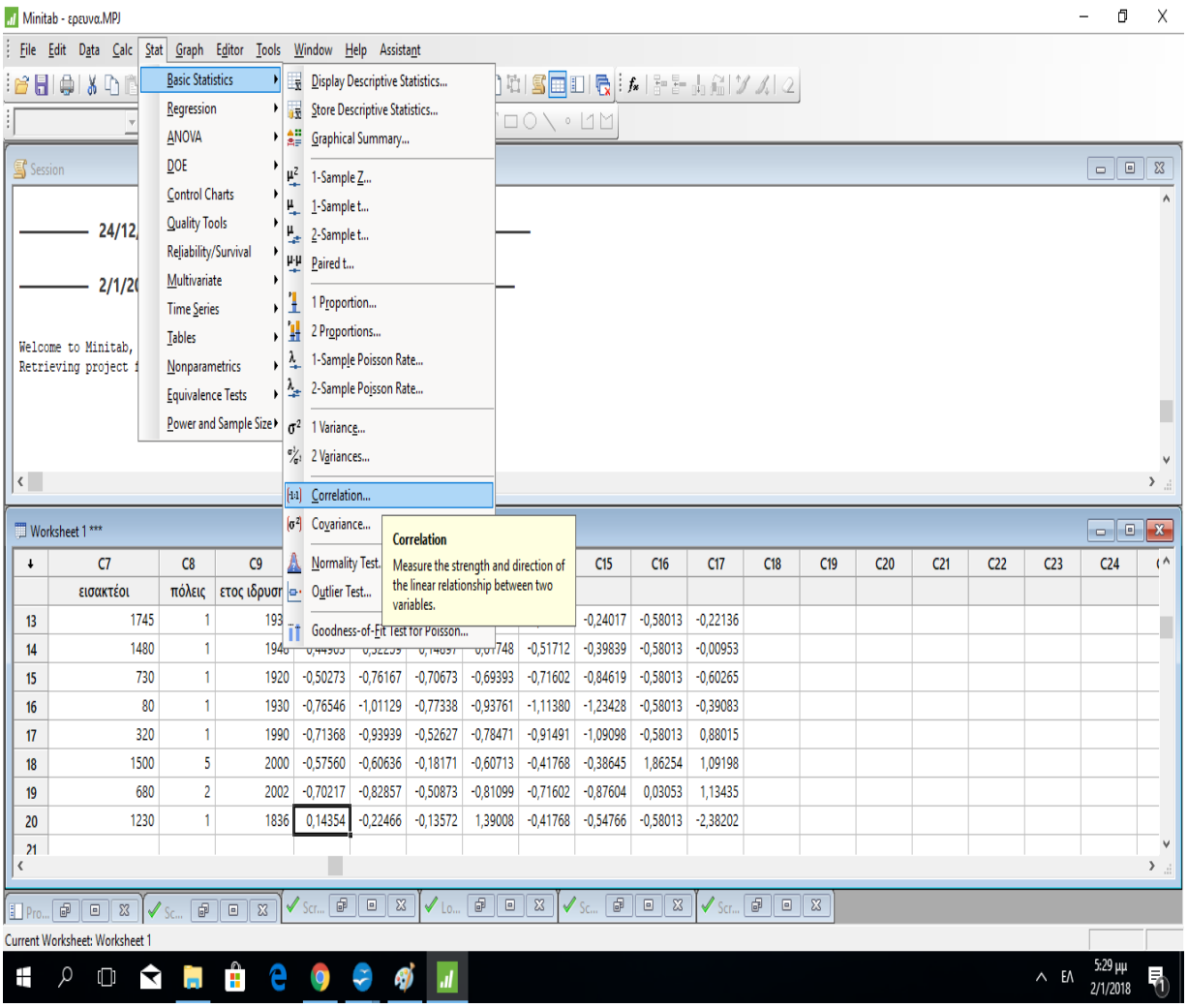
Παρατηρούμε ότι έχουμε και θετικές και αρνητικές συσχετίσεις. Μερικές από τις συσχετίσεις είναι πολύ ισχυρές. Π.χ. από την $1^{η}$ στήλη του πίνακα βλέπουμε ότι η μεταβλητή ΜΕΛΗ ΔΕΠ σχετίζεται ισχυρά με τις μεταβλητές Προπτυχιακοί, Διδακτορικοί, Τμήματα και Εισακτέοι (συντελεστές συσχέτισης 0,97, 0,92, 0,94, 0,90 αντίστοιχα), οπότε έχει νόημα η Α.Κ.Σ.

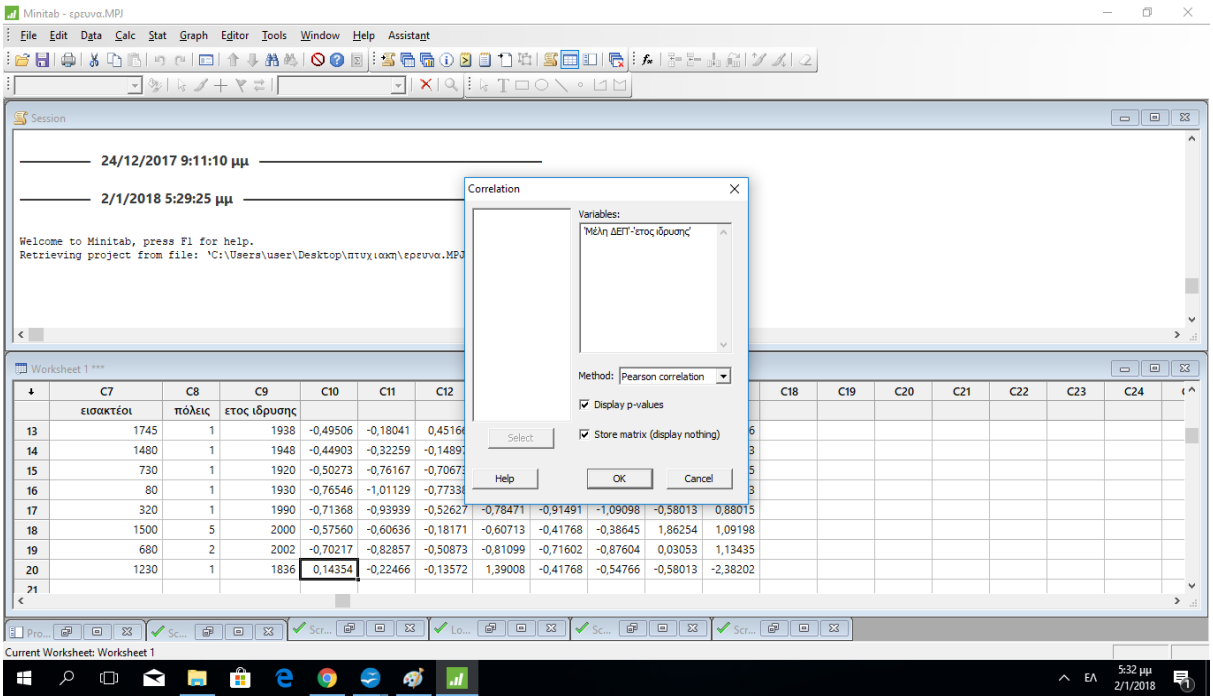
Στη συνέχεια θα υπολογίσουμε τις **ιδιοτιμές** του πίνακα συσχετίσεων.

ΒΗΜΑΤΑ στο Minitab

Αρχικά θα δημιουργήσουμε τον πίνακα (συν)διακύμανσης του κανονικοποιημένου πίνακα των δεδομένων στο Minitab ακολουθώντας τις εντολές:

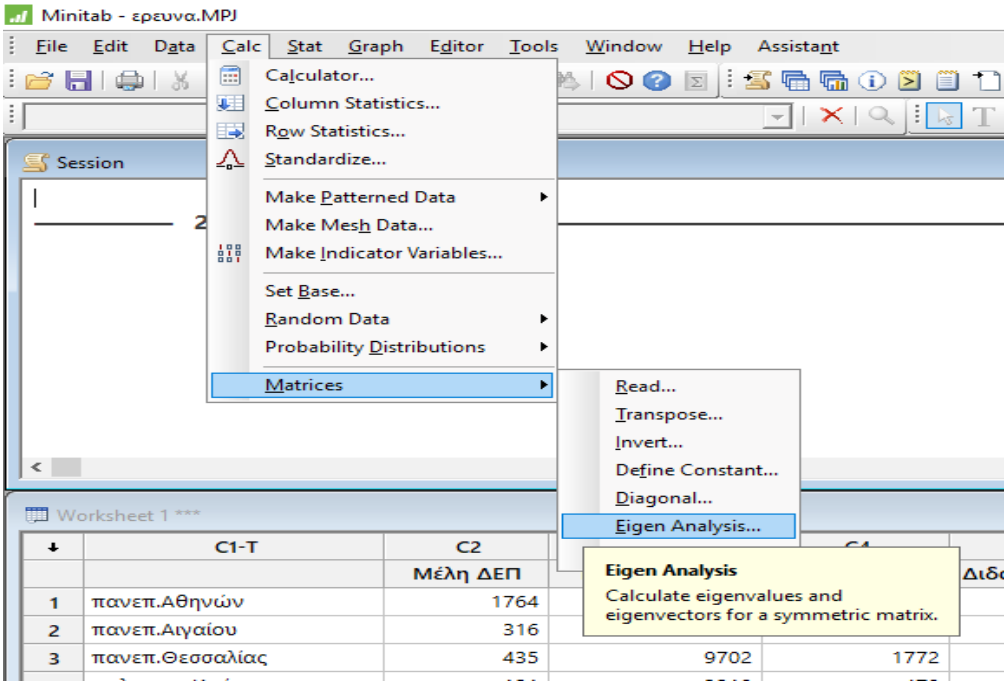
STAT →BASIC STATISTICS →CORRELATION→επιλογή και των 8 μεταβλητών

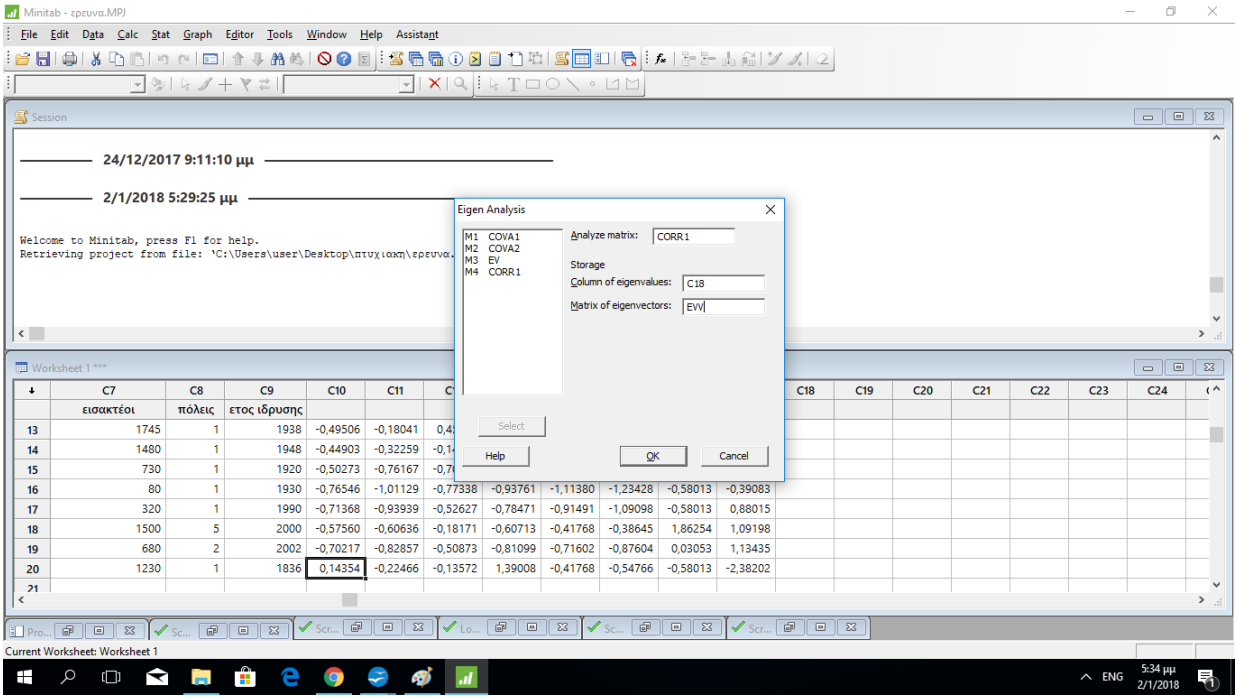




Στη συνέχεια θα υπολογίσουμε τις ιδιοτιμές του.

CALC → MATRICES →EIGEN ANALYSIS→ στο παράθυρο analyze matrix επιλογή πίνακα CORR1(για υπολογισμό ιδιοτιμών), column of eigenvalues: C18 (στήλη που θέλω να εμφανιστούν ιδιοτιμές), matrix of eigenvectors: EVV (ονομασία πίνακα ιδιοδιανυσμάτων).





Και παίρνουμε

Worksheet 1 ***

	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
	τμήματα	εισακτέοι	πόλεις	ετος ιδρυσης									ιδιοτιμές
1	33	5455	1	1837	2,54069	2,29541	3,76432	2,50973	1,96905	1,97493	-0,58013	-2,36084	5,64223
2	17	2855	6	1984	-0,23617	-0,05368	-0,21757	-0,48051	0,37790	0,42257	2,47320	0,75305	1,50224
3	18	2510	5	1984	-0,00796	0,01788	-0,14117	-0,03217	0,47734	0,21658	1,86254	0,75305	0,48048
4	5	720	1	1984	-0,61012	-0,73518	-0,64515	-0,73056	-0,81546	-0,85216	-0,58013	0,75305	0,23789
5	41	6000	1	1925	2,88588	2,68930	1,29863	2,12749	2,76462	2,30033	-0,58013	-0,49674	0,08670
6	19	3835	4	1973	0,26819	0,56289	0,18818	0,49739	0,57679	1,00769	1,25187	0,52004	0,04499
7	6	825	1	1984	-0,62930	-0,78153	-0,66074	-0,58801	-0,71602	-0,78947	-0,58013	0,75305	0,00345
8	15	2930	1	1970	0,13587	0,20664	-0,23979	0,47589	0,17900	0,46735	-0,58013	0,45649	0,00202
9	16	2710	2	1973	0,08793	0,20048	-0,30527	-0,03217	0,27845	0,33600	0,03053	0,52004	

Οι 8 ιδιοτιμές του πίνακα είναι οι (προσεγγιστικά):

$$\lambda_1 = 5,64, \quad \lambda_2 = 1,50, \quad \lambda_3 = 0,48, \quad \lambda_4 = 0,23,$$

$$\lambda_5 = 0,006, \ \lambda_6 = 0,004, \ \lambda_7 = 0,003, \ \lambda_8 = 0,002$$

και εμφανίζονται στη στήλη C18. Από τα αντίστοιχα ιδιοδιανύσματα που εμφανίζονται στις στήλες C10-C17, μόνον ορισμένα από αυτά παρέχουν σημαντική πληροφορία. Πρόκειται για τις κύριες συνιστώσες (συντελεστές).

Σαν κριτήριο επιλογής των κυρίων συνιστωσών θα χρησιμοποιήσουμε αυτό που αναφέρεται στο *ποσοστό της συνολικής διακύμανσης που ερμηνεύεται από τις κύριες*

συνιστώσες. Θα κρατήσουμε λοιπόν τόσους παράγοντες/συνιστώσες, όσοι αθροιστικά ερμηνεύουν το 70-80% της συνολικής διακύμανσης των δεδομένων.

Ξεκινώντας με την πρώτη και συγχρόνως μεγαλύτερη ιδιοτιμή (οι ιδιοτιμές εμφανίζονται σε φθίνουσα διάταξη μεγέθους) λ_1=5,64223 που αντιστοιχεί στην πρώτη κύρια συνιστώσα, υπολογίζουμε το ποσοστό ολικής διακύμανσης-αδράνειας που ερμηνεύεται από τον πρώτο κύριο άξονα /κύρια συνιστώσα:

$$\frac{\lambda_1}{\sum_{i=1}^{8}\lambda_i} = \frac{5,64223}{8} = 0,705 \quad (\text{δηλαδή } \mathbf{70.5\%})$$

Το ποσοστό ολικής αδράνειας/διακύμανσης που ερμηνεύεται από το δεύτερο κύριο άξονα/κύρια συνιστώσα είναι ίσο με:

$$\frac{\lambda_2}{\sum_{i=1}^{8}\lambda_i} = \frac{1,5}{8} = 0,188 \ (\mathbf{18.8\%})$$

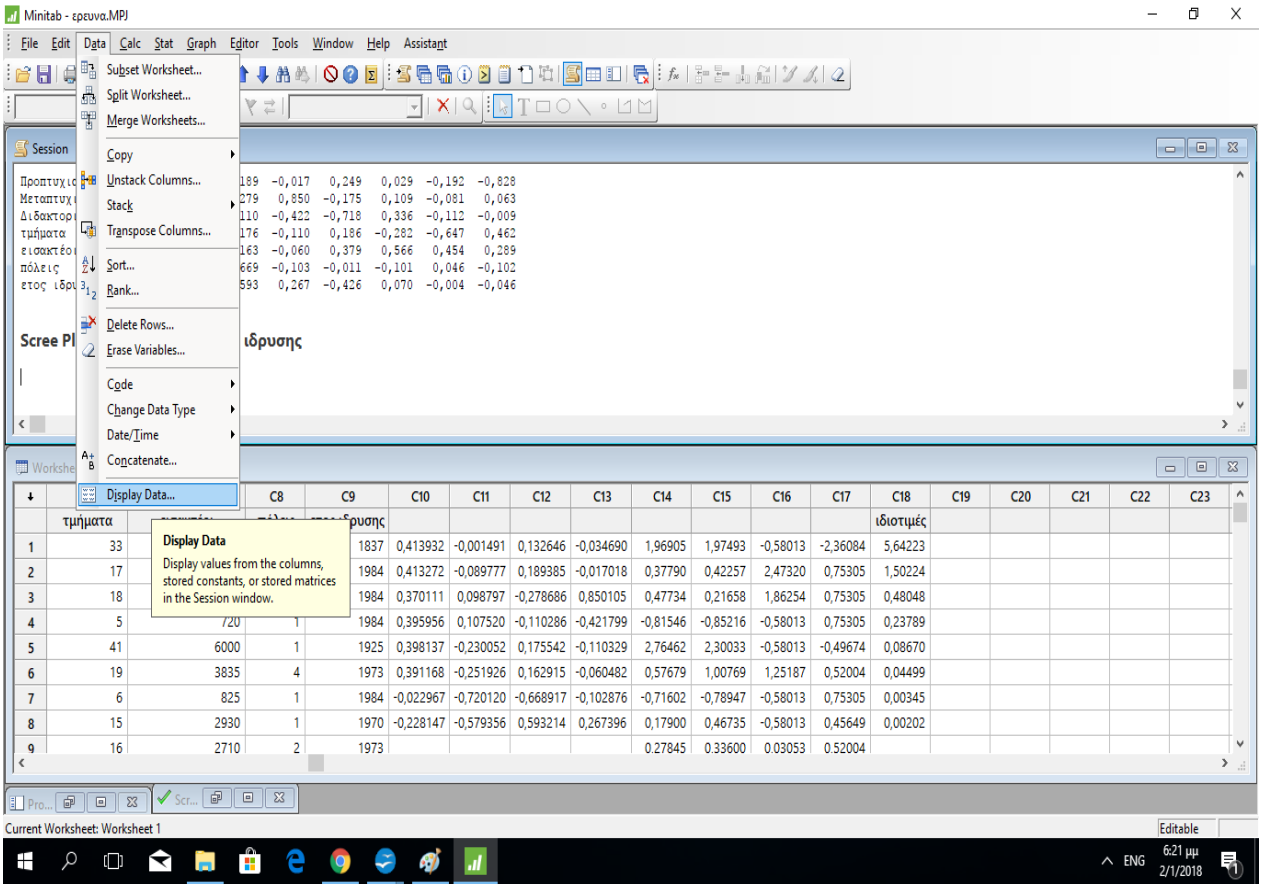
Το επίπεδο λοιπόν που δημιουργείται από τους δύο πρώτους άξονες/κύριες συνιστώσες ερμηνεύει το 89,3% της ολικής αδράνειας/διακύμανσης.

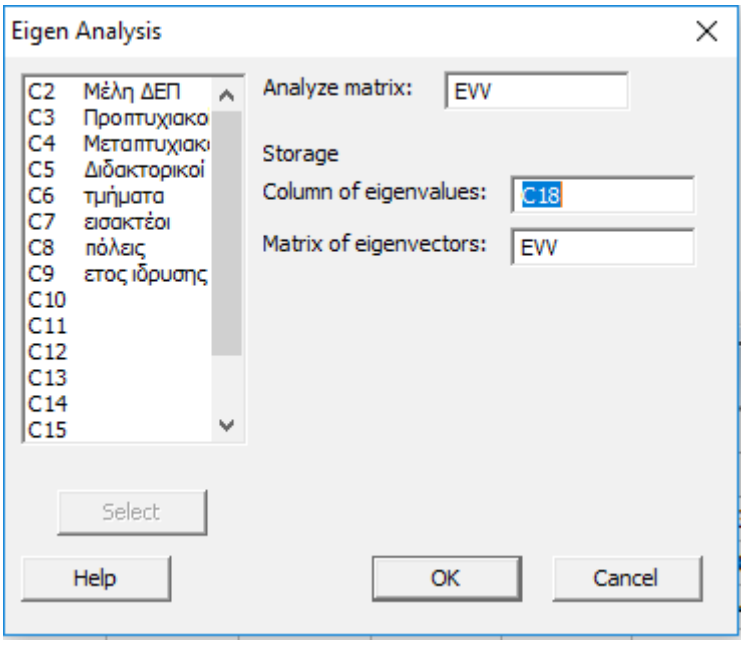
Αυτό το 89,3% της ολικής αδράνειας/διακύμανσης μας επιτρέπει να συμπεράνουμε ότι το νέφος των σημείων στον χώρο των 8 διαστάσεων είναι πολύ πεπλατυσμένο/ διεσπαρμένο και (με σφάλμα 10,7%) μπορούμε να το θεωρήσουμε επίπεδο. Η απεικόνιση του λοιπόν στο επίπεδο των δύο πρώτων κύριων αξόνων (αξόνων αδράνειας) θα είναι πολύ ικανοποιητική.

Θα υπολογίσουμε (με τη βοήθεια του minitab) τα ιδιοδιανύσματα που αντιστοιχούν στον πρώτο και στον δεύτερο κύριο άξονα (ιδιοδιανύσματα που αντιστοιχούν στις 2 πρώτες ιδιοτιμές).

Για τον υπολογισμό των ιδιοδιανυσμάτων ακολουθούμε τις παρακάτω εντολές:

Data →Display Data →column of eigenvalues: C18





Τα ιδιοδιανύσματα εμφανίζονται στον παρακάτω πίνακα:

Matrix EVV

0,413932	-0,001491	0,132646	-0,034690	-0,177206	-0,678266	0,563473	0,030189
0,413272	-0,089777	0,189385	-0,017018	0,249177	0,028776	-0,192111	-0,827748
0,370111	0,098797	-0,278686	0,850105	-0,174822	0,109047	-0,080763	0,062740
0,395956	0,107520	-0,110286	-0,421799	-0,718418	0,336064	-0,111505	-0,009236
0,398137	-0,230052	0,175542	-0,110329	0,185509	-0,281631	-0,646633	0,462291
0,391168	-0,251926	0,162915	-0,060482	0,378712	0,565529	0,454241	0,289380
-0,022967	-0,720120	-0,668917	-0,102876	-0,010595	-0,101260	0,046332	-0,101756
-0,228147	-0,579356	0,593214	0,267396	-0,426467	0,070030	-0,004490	-0,045746

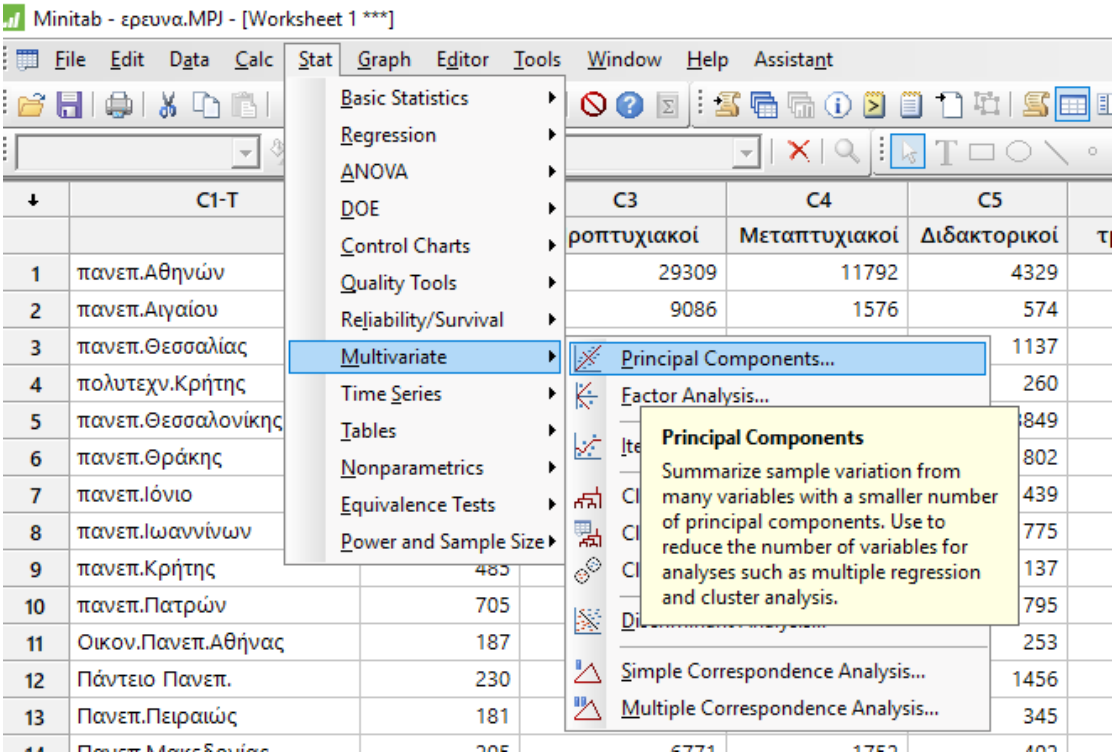
Για να υπολογίσουμε τις συντεταγμένες των παρατηρήσεων, ως προς τις (δύο) κύριες συνιστώσες, πολλαπλασιάζουμε τον 20x8 πίνακα δεδομένων με τον 8x2 πίνακα των διευθύνσεων των κύριων συνιστωσών (πίνακας με πρώτη και $2^{η}$ στήλη τα ιδιο-διανύσματα που αντιστοιχούν στην $1^{η}$ και $2^{η}$ ιδιοτιμή του πίνακα των συσχετίσεων).

6,4956827722	1,266898716
-0,3035833684	-2,478670094
-0,0007095965	-1,960912207
-1,9009103164	0,308383583
5,7561717247	-0,398655886
1,0866568372	-1,568203825
-1,8132152212	0,287687569
0,404579311	0,003100302
0,1164725539	-0,523729368
1,6738941224	-0,836764331
-0,8829568541	0,926422379
-0,4917918166	0,883463799
-0,5711909994	0,692891126
-0,9650396013	0,59113984
-1,5244609601	1,069511074
-2,2160405676	1,126085429
-2,1676228728	0,342203172
-1,4058694216	-1,808385903
-2,029730356	-0,355784158
0,6430081444	2,187869511

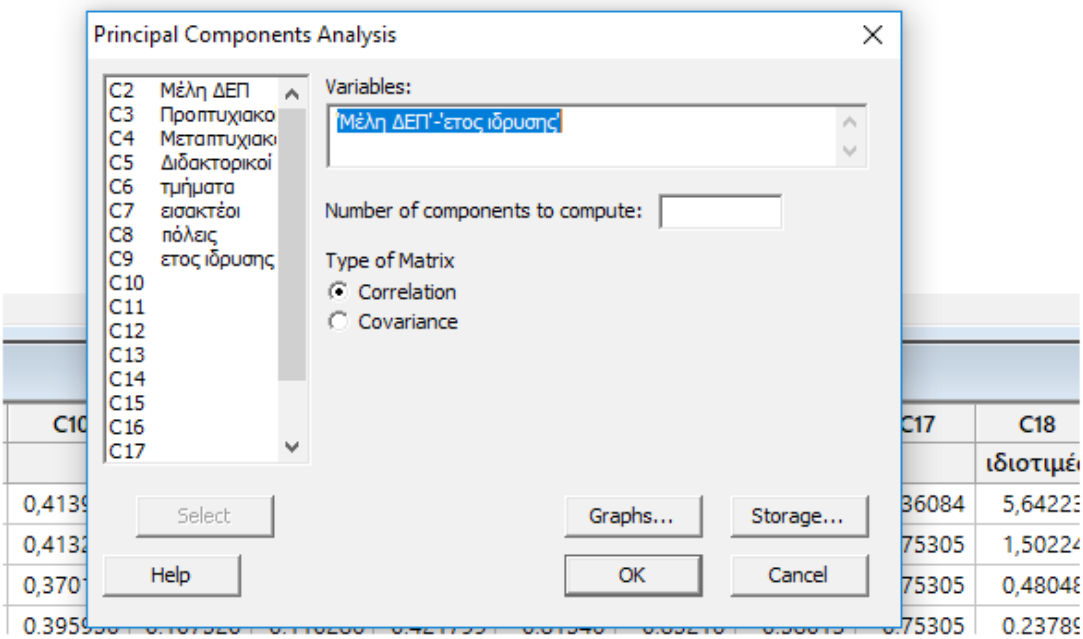
Συνεχίζοντας, θα επαληθεύσουμε τα παραπάνω αποτελέσματα καθώς και θα αναλύσουμε/ερμηνεύσουμε τα γραφήματα που προκύπτουν από την Ανάλυση Κυρίων Συνιστωσών

Στο στο minitab χρησιμοποιούμε τις εντολές:

STAT → MULTIVARIATE ->PRINCIPAL COMPONETS



Επιλέγω ως variables κάθε μια από τις 8 μεταβλητές μας και επίσης επιλέγω όλα τα γραφήματα. Ακόμα αποφασίζουμε να χρησιμοποιήσουμε τον πίνακα συσχετίσεων (correlation matrix)



και έχουμε...

Principal Component Analysis: Μέλη ΔΕΠ; Προπτυχιακοί; Μεταπτυχιακο; Διδακτορικοί;

Eigenanalysis of the Correlation Matrix

Eigenvalue	5,6422	1,5022	0,4805	0,2379	0,0867	0,0450	0,0035	0,0020
Proportion	0,705	0,188	0,060	0,030	0,011	0,006	0,000	0,000
Cumulative	0,705	0,893	0,953	0,983	0,994	0,999	1,000	1,000

Στην πρώτη γραμμή του πίνακα βλέπουμε τις ιδιοτιμές του πίνακα των συσχετίσεων, στη δεύτερη γραμμή παρατηρούμε το ποσοστό της ολικής αδράνειας που ερμηνεύεται από την κάθε κύρια συνιστώσα. Στην τρίτη γραμμή βλέπουμε αθροιστικά το ποσοστό της ολικής αδράνειας που ερμηνεύεται από αυτή και όλες τις προηγούμενες συνιστώσες.

Επίσης εμφανίζονται, σαν στήλες, και τα ιδιοδιανυσματα / κύριες συνιστώσες (συντελεστές)

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Μέλη ΔΕΠ	0,414	-0,001	0,133	-0,035	-0,177	-0,678	0,563	0,030
Προπτυχιακοί	0,413	-0,090	0,189	-0,017	0,249	0,029	-0,192	-0,828
Μεταπτυχιακοί	0,370	0,099	-0,279	0,850	-0,175	0,109	-0,081	0,063
Διδακτορικοί	0,396	0,108	-0,110	-0,422	-0,718	0,336	-0,112	-0,009
τμήματα	0,398	-0,230	0,176	-0,110	0,186	-0,282	-0,647	0,462
εισακτέοι	0,391	-0,252	0,163	-0,060	0,379	0,566	0,454	0,289
πόλεις	-0,023	-0,720	-0,669	-0,103	-0,011	-0,101	0,046	-0,102
ετος ιδρυσης	-0,228	-0,579	0,593	0,267	-0,426	0,070	-0,004	-0,046

Και εμείς επιλέγουμε όπως είδαμε προηγουμένως τις στήλες PC1 και PC2 που αντιστοιχούν στις διευθύνσεις των κύριων συνιστωσών και ερμηνεύουν το αντιστοιχούν στο 89,3 % της ολικής αδράνειας/ διακύμανσης των δεδομένων.

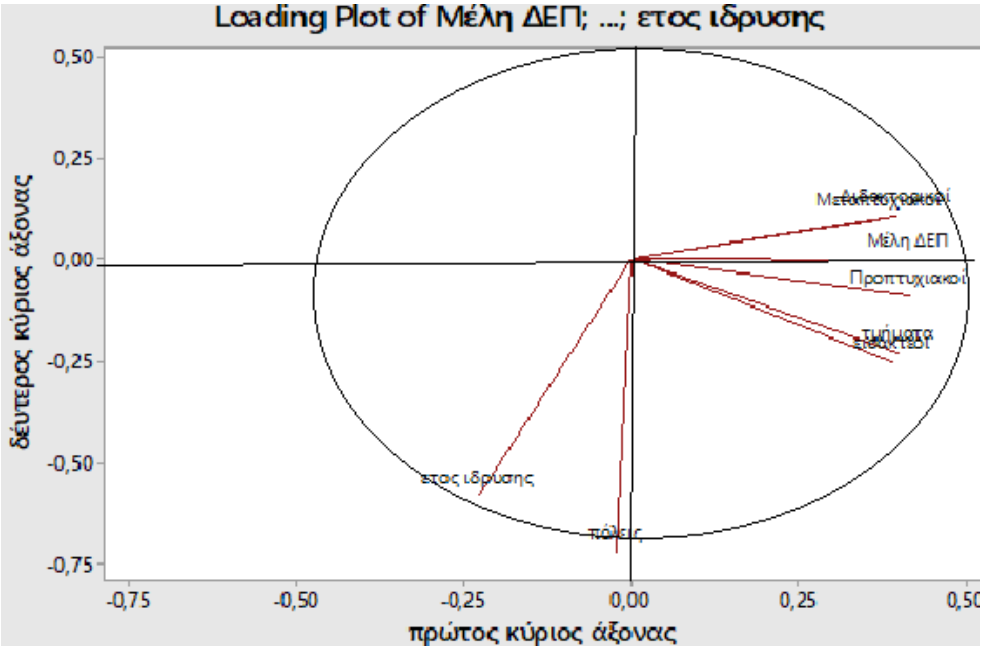
Η συντεταγμένη μιάς παρατήρησης ως προς την $1^{η}$ κύρια συνιστώσα δίνεται από τη σχέση:

$$Y_1 = 0,414(\text{Μέλη ΔΕΠ}) + 0,413(\text{Προπτυχιακοί}) + 0,370(\text{Μεταπτυχιακοί}) + 0,396(\text{Διδακτορικοί}) + 0,398(\text{Τμήματα}) + 0,391(\text{εισακτέοι}) + 0,023(\text{πόλεις}) + 0,228(\text{έτος ίδρυσης})$$

Ανάλογα ορίζεται η συντεταγμένη μιάς παρατήρησης ως προς τη $2^{η}$ κύρια συνιστώσα

ΕΡΜΗΝΕΙΑ ΓΡΑΦΗΜΑΤΩΝ

Παρουσιάζουμε παρακάτω τα γραφήματα που παράγει το minitab, και επίσης δίνουμε και την ερμηνεία καθενός από αυτά.

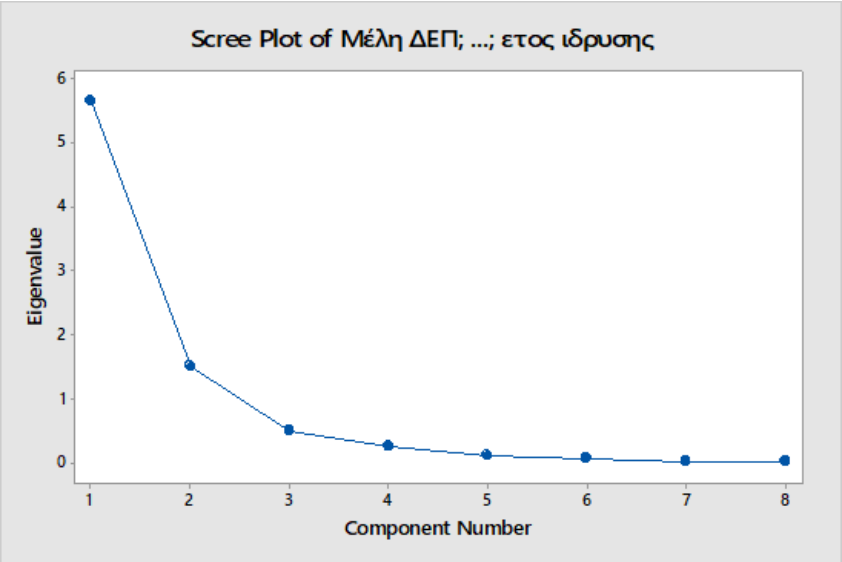


Loading plot: κύκλος συσχετίσεων της ΑΚΣ

Ο παραπάνω κύκλος συσχετίσεων μας δείχνει ότι δυό μεταβλητές που βρίσκονται κοντά η μια στην άλλη επί της περιφέρειας είναι ισχυρά συσχετισμένες μεταξύ τους.

Επίσης παρουσιάζονται οι κατευθύνσεις προς τις οποίες κάθε μεταβλητή αυξάνει.

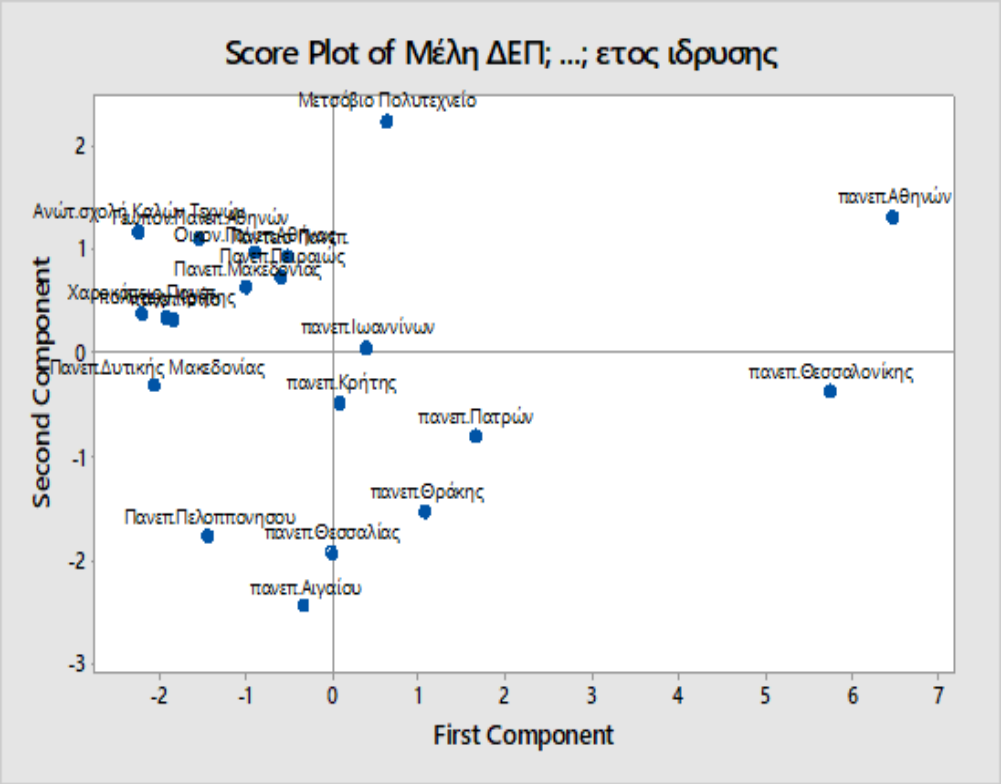
Π.χ η μεταβλητή Μεταπτυχιακοί είναι έντονα συσχετισμένη με τη μεταβλητή Διδακτορικοί ενώ είναι ασυσχέτιστη με την μεταβλητή έτος ίδρυσης. Αντίστοιχα η μεταβλητή Μέλη ΔΕΠ είναι συσχετισμένη με την μεταβλητή Προπτυχιακοί ενώ είναι ασυσχέτιστη με την μεταβλητή έτος ίδρυσης και τη μεταβλητή πόλεις.



Scree plot: διάγραμμα που περιγράφει τη σχέση μεταξύ του αύξοντα αριθμού συνιστωσών και την αντίστοιχη τιμή των ιδιοτιμών.

Σύμφωνα με τον Cattel οι συνιστώσες που αντιστοιχούν σε μικρή λ_j και ταυτόχρονα δεν διαφέρουν μεταξύ τους σημαντικά $\lambda_j \approx \lambda_{j+1}$ δεν αποτελούν κυρίες συνιστώσες. Επομένως από το διάγραμμα συμπεραίνουμε ότι μόνο οι 2 πρώτες ιδιοτιμές θα γίνουν δεκτές.

Εάν χρησιμοποιήσουμε το κριτήριο **scree plot (**διαλέγουμε τόσες συνιστώσες όσος και ο αριθμός των j, ώσπου η γραμμή που συνδέει τα $\left(j, \lambda_j\right)$, $j = 1, 2, ..., p$ να γίνει οριζόντια) τότε βλέπουμε ότι από το τρίτο σημείο η γραμμή τείνει να γίνει οριζόντια, επομένως και αυτό είναι μια ένδειξη ότι δύο κύριες συνιστώσες ίσως είναι αρκετές.



score plot: στο παραπάνω γράφημα παριστάνονται οι παρατηρήσεις (συντεταγμένες τους) ως προς τις δύο νέες κύριες συνιστώσες

Τα σημεία που είναι γειτονικά μεταξύ τους προσδιορίζουν υψηλό βαθμό ομοιότητας και όσο περισσότερο απέχουν από την αρχή των αξόνων τόσο ισχυρότερη είναι η δράση που ασκούν στην ανάλυση κυρίων συνιστωσών.

Παρατηρούμε λοιπόν ότι το Πολυτεχνείο Κρήτης παρουσιάζει πολλές ομοιότητες με το Χαροκόπειο Πανεπιστήμιο, το Οικονομικό Πανεπιστήμιο, το Πανεπιστήμιο Μακεδονίας, το Πανεπιστήμιο Πειραιώς, ενώ κανένα από τα ΑΕΙ που μελετάμε δεν παρουσιάζει ομοιότητες με το Πανεπιστήμιο Αθηνών και το Πανεπιστήμιο Θεσσαλονίκης. Μια μικρή ομοιότητα παρουσιάζεται ανάμεσα στο πανεπιστήμιο Αθηνών και το Πανεπιστήμιο Θεσσαλονίκης.

Κάτι άλλο που παρατηρούμε είναι την υψηλή συνεισφορά του Πανεπιστημίου Αθηνών στην αδράνεια του πρώτου άξονα (65%) και την υψηλή συνεισφορά του Πανεπιστημίου Αιγαίου στην αδράνεια του δεύτερου άξονα (24,66%).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Βιβλία

- Γναρδέλλης Χ, *Εφαρμοσμένη Στατιστική*, Εκδ. Παπαζήση, 2003.
- Καρλής Δ., *Πολυμεταβλητή Στατιστική Ανάλυση*, 2005, Σταμούλης.
- Μαυρομάτης Γ., *Στατιστικά μοντέλα και μέθοδοι ανάλυσης δεδομένων*, 1999, University studio press, Θεσσαλονίκη.
- Μπεχράκης Θ., *Πολυδιάστατη Ανάλυση Δεδομένων*, 1999, Λιβάνη.
- Μπεχράκης Θ., *Στατιστική για τις επιστήμες του ανθρώπου και της κοινωνίας*, 2010, Λιβάνη.
- Ντζούφρας Ι, *Στοιχεία Πολυμεταβλητής Ανάλυσης Δεδομένων*, 2002, Πανεπιστημιακές παραδόσεις Πανεπιστήμιο Αιγαίου.
- Παπαδημητρίου Γ., *Η Ανάλυση δεδομένων*, 2007, Εκδ. Δαρδανός
- Σιάρδος Γ., *Μέθοδοι πολυμεταβλητής Στατιστικής Ανάλυσης*. Εκδ. Σταμούλης, 2006, 3η έκδοση

Ιστοσελίδες

- https://repository.kallipos.gr/bitstream/11419/2129/1/05_chapter04.pdf
- https://eclass.uoa.gr/modules/document/file.php/DI367/%CE%A5%CE%BB%CE%B9%CE%BA%CF%8C/PCA_method.pdf
- https://www.youtube.com/watch?v=SaEmG4wcFfg
- https://www.youtube.com/watch?v=Ao_iYZ50RNY